Vol.2013-MUS-100 No.24
2013/9/1

スペクトルモーフィングによるグロウル系統の歌唱音声合成

Bonada Jordi[1,a)]　Blaauw Merlijn[1,b)]　才野 慶二郎[2,c)]　久湊 裕司[2,d)]

概要：音声，特に強め表現を伴う歌声においてはしばしば声帯振動に基本周期外の不規則な挙動が見られる．本稿ではこのような声質を持つ音声を再現するためのスペクトルモーフィングに基づいた音声合成手法について述べる．本手法は，ターゲットとなる声質を持った音声サンプルの励振源に相当する成分と，入力音声のスペクトル包絡を用いて合成を行うものである．まず，声質ターゲットサンプルに対し，その基本周波数を入力音声の基本周波数に合わせこむための時間領域リサンプリング処理を行う．その後，声質ターゲットサンプルのスペクトルの元々の包絡構造をなるべく復元するように，調波成分の再配置を行う．最後に，そこに入力音声の調波の振幅と位相を適用することで，入力音声の音色と声質ターゲットサンプルの声質を併せ持つ音声信号を得る．その音声信号と入力音声を任意の比率でモーフィングすることで，声質ターゲットサンプルの声質を任意の分量だけ持つ音声が合成可能となる．本稿では，グロウルの声質を持つ音声を使用した歌声合成および主観評価実験を行った．

Bonada Jordi[1,a)]　Blaauw Merlijn[1,b)]　Saino Keijiro[2,c)]　Hisaminato Yuji[2,d)]

Abstract: In this paper we introduce a morph-based approach for generating voice source aperiodicities frequently associated with strong vocal expressions, especially in singing. In our approach the excitation characteristics of one signal are combined with the fundamental frequency and spectral envelope characteristics of another signal. An exemplar sustained sample of the target voice quality is resampled in the time domain in order to generate a continuous signal matching the input voice's fundamental frequency. While we found the temporal scaling to be acceptable in many contexts, the frequency scaling has to be inverted in order to generate appropriate spectral content for the source excitation's entire bandwidth. Finally, the input signal's harmonic amplitudes and phases are applied to the transformed morph sample, allowing for a simple one-dimensional control of morph amount by linear interpolation with the input signal. The proposed system is evaluated and the results are discussed.

1. はじめに

歌唱において，歌い手が自身の歌声に込める歌唱表現は，楽曲のニュアンスや歌い手自身の情動などを伝え，そのことが歌声そのものをより魅力的なものにする．コンピュータ上で歌声を人工的に生成する歌声合成技術において，合成音声にそういった表現力を持たせられるようにすることは，重要な課題である．

話し声・歌声を問わず，音声は，声帯振動に伴い発生する音源（励振源）が，声道通過時にその形状に影響を受けて調音（フィルタリング）されて，口腔・鼻腔外へと放出していくことで形成される．本研究では，音源部の特徴が“声質 (voice quality)” に，調音部の特徴が “音色 (timbre)” に表れるものと考える[*1]．なお，無声音の場合は声帯振動を伴わないため上記とは異なるが，本稿においては声質の知覚は有声音部分が支配的であると考え，以降，有声音のみを対象とした議論を行う．

話し言葉に関する先行研究で，音声に多様な表現をもたらすためには，基本周波数やパワーなど韻律的特徴の他に声質が重要な役割を持つことが報告されている [1,2]．歌声でも同様に，たとえば，ジャズにおいて Louis Armstrong がしばしば行うような荒々しい歌唱法や，日本の演歌において “うなり” と呼ばれる歌唱法に現れるような声質は，

[1] Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
[2] Development Department 1, Research & Development Division, Yamaha Corporation
a) jordi.bonada@upf.edu
b) merlijn.blaauw@upf.edu
c) keijiro_saino@gmx.yamaha.com
d) yuji_hisaminato@gmx.yamaha.com

[*1] “声質” という用語は，声道形状に起因する話者性のような，音声が伝える情報の中で言語情報と韻律を除いた成分を指す広義で使用されることも多いが，本稿では，声道形状に非依存な音源部分の発声様式を指す狭義で使用するものとする．

通常発声 (modal な声質) の歌唱とは大きく異なった印象を与え，そのことが歌声に激しさのような表現を効果的に与える．この声質は，声帯上部に位置する披裂喉頭蓋ひだが声帯振動の基本周期よりも長い周期で声帯と共に強く振動することによりもたらされることがわかっている [3]．この声質を，本研究ではグロウルと呼ぶこととする[*2]．

グロウルの特筆的な音響特徴として，基本周期および振幅の揺らぎの指標である jitter および shimmer が大きい値を示す，基本周期 T による周期構造の他に $2T$〜$3T$ 程度のマクロピリオドにより成されるマクロパルスと呼ばれる周期構造を持つ，スペクトル上で各調波の間にサブハーモニクスと呼ばれるピーク成分を持つ，などがある [4]．一方，これらの特徴を有しながら，harsh, hoarse, rough などと呼ばれグロウル (growl) と区別される声質がある．これらの声質は上記のような特徴がそれぞれ異なった形で現れることにより，それぞれがそのいずれとも異なる聴感をもたらしていると考えられる．本研究ではそれらを総称して，“グロウル系統の声質” と呼ぶこととする．グロウル系統の声質はジャンルを問わず歌唱の強め表現の際にしばしば用いられる．しかしながら，modal な声質と比較してその特徴が特異であるため，従来の実用的な歌声合成システムにおいては例外的な存在として扱われ，その再現にはなかなか至っていなかった．そこで本研究では，実用的な歌声合成システムにおいてグロウル系統の声質を持った歌声を合成するための手法について述べる．以下，本稿ではグロウル系統の声質を持った歌声（音声）をグロウル系統の歌声（音声）と呼ぶ．

以下，2 章で関連する先行研究についてまとめ，3 章で本研究で提案するスペクトルモーフィングによるグロウル系統の歌声の合成手法について述べる．4 章で歌声合成実験，主観評価結果の記述およびその考察を行い，最後に 5 章で全体のまとめと今後の課題を示す．

2. 関連研究

グロウル系統の音声の合成を実現するするためのアプローチは，大きく分けてモデルベースとサンプルベースの 2 つの手法が考えられる．それぞれのアプローチの関連研究についてまとめる．

2.1 モデルベースの手法

モデルベースの手法は，もともと modal な声質の音声に，何らかのモデルで表現されるパラメトリックな変動を付与することで，グロウル系統の声質を再現しようというものである．Schoentgen らは，音声中の jitter や shimmer を表現するための様々なモデルを紹介・提案した [6]．パラメタライズされた変調は，例えば TD-PSOLA などの手法により音声の励振源パルスに対して付与されることで分析元の音声が持つ jitter や shimmer の再現が可能となる．[7], [8] ではそのような手法を用いて声質変換を行う応用が提案されている．また，Loscos らはサブハーモニクスに着目し，その時間周波数領域における振幅と位相の挙動をモデル化した [9]．

このアプローチはの難しさは，現象を適切にモデル化することそのものにある．特にグロウル系統の声質は，音源生成機構の不規則さによりもたらされるものであり，その不規則さを適切に追従できるモデルを設計することは，決して容易なことではない．また，単純なモデルでは，グロウル系統の声質の中でもグロウルとそれ以外の声質の差異を表現できないという懸念も生まれる．

2.2 サンプルベースの手法

サンプルベースの手法は，所望の声質を持った自然音声をあらかじめ用意しておき，そのサンプルを直接用いた音声合成を行うというものである．モデルベースの手法と比較すると，こちらの手法を用いた non modal な声質の音声合成に取り組む先行研究はそれほど多くはない．

音声の信号を音源部と調音部に分解するなど適切な特徴量へ変換した上で，発話内容が同一かつ時間同期のとれたパラレルレコーディングの間で特徴量をモーフィングするような応用は数多く提案されてきたが [10–12]，それらのほぼすべてが modal な声質の音声を前提としている．

これは，世界中で広く利用されている音声分析合成系の STRAIGHT [13] を含め，音声の分析・変換・再合成に利用されるボコーダのほとんどが有声音に関して基本周波数の整数倍成分のみの調波構造を仮定していて，サブハーモニクスや jitter，shimmer の表現に関して不十分な面があるためである．近年では徐々に，励振源に関する特徴を捉えることで音声合成の品質向上を目指すような研究が増えつつあるが [14, 15]，サブハーモニクスなどの特徴は仮定せず声質のハリ成分[*3]を表現するに留まる場合が多い．Kawahara ら [16] はさらに踏み込んで，グロウル系統の音声で観測される高速な基本周波数の変動を追従できるような瞬時周波数の表現を導入し，実際にグロウル系統の音声の分析再合成の結果，その声質が再現されたことを確認している．ただし同時に，補間や変形後の合成品質については未だ研究途上であるとも述べている．

また，根本的な問題として，ユーザの任意の曲を歌わせる実用的な歌声合成システムを目指す場合，パラレルレコー

[*2] 近年では，一般にデスボイスと呼ばれるような音声や，シャウト，スクリームと呼ばれるような音声を対象とした分析の研究も散見されるようになってきたが [4, 5]，それらの声質の呼称については定まっていないのが現状である．例えば [4] では本稿でグロウルと呼ぶ声質をポップグロウルと定め，日本で一般にデスボイスとカテゴライズされる唸り発声と区別している．本稿では [3] で growl と表現されている声質をそのままグロウルと呼称する．

[*3] [14] では lax, modal, tense と表現されている．

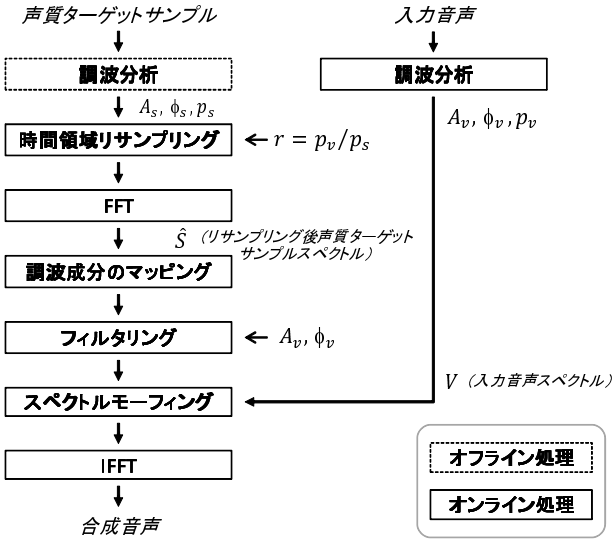

図 1 提案手法のブロック図（1 フレームに関する処理）
Fig. 1 Block diagram of the proposed system.

ディングが必要という事実は極めて大きなネックとなる．

3. スペクトルモーフィングによるグロウル系統の歌声の合成

本章では，本稿で提案するスペクトルモーフィングによるグロウル系統の歌声の合成手法について述べる．本研究で目標とするのは，なるべく実用的な歌声合成システムにおいて，グロウル系統の声質を持った歌声の合成を可能にすることである．そこで満たすべき要件を，以下のように定める．

(1) 合成音声において，グロウル系統の声質がなるべく高い自然性で再現される
(2) パラレルレコーディングをあらかじめ用意しておくことを要求しない
(3) 歌唱内の任意の箇所に，任意の量だけ声質成分の付与が可能である

3.1 手法の流れ

3.1.1 フレームバイフレーム処理の進行

まず，声質ターゲットサンプルとして，グロウル系統の声質を持ちその声質や韻律が安定した持続発声のサンプルをあらかじめ用意し，その基本周波数 p_s の抽出およびスペクトル中の各調波の振幅 A_s, 位相 ϕ_s の計算を行っておく．また，合成処理のループの始点と終点も人手で決定しておく．システムに任意の音声が入力されると，その先頭から順にフレームバイフレームで処理を行うことになるが，並行して声質ターゲットサンプルもフレームバイフレーム処理を進行させる．ただし通常，声質ターゲットサンプルの継続長は入力音声よりずっと短いので，ループの始点から終点までの間を，入力音声が終了するまでの間反復的に往復するようにフレーム進行させることで，任意の時間長の合成処理を実現する．ただし，始点と終点を固定してしまうと，指定区間長を周期とする人工的な揺らぎが知覚されてしまうので，処理方向を反転させるたびにその端点にランダムなオフセットを加える．

ところで，グロウル系統の声質をモデルベースで再現するためのネックの 1 つに，音声中の揺らぎの長時間的な挙動の再現が難しいということがある．例えば 1 フレームにおける，サブハーモニクスを含むスペクトル形状をモデルにより表現できたとしても，時間経過に伴うその形状の変化までも適切に表現できなければ，自然には聞こえない．その点において本手法は，声質の静的な特徴だけでなく動的な特徴も含めてサンプルを写実的に利用することで再現できるという大きなメリットを持つ．

3.1.2 1 フレームの合成処理

フレームバイフレーム処理中のある時刻において，声質ターゲットサンプルと入力音声の各フレームを使用して，1 フレームの合成処理を行う．ただし，この処理は入力音声の処理フレームが有声のときのみ行い，無声フレームの場合は何も処理せずそれをそのまま合成フレームとする．処理の概要を図 1 に示す．

【調波分析】
入力音声に対して，基本周波数 p_v の抽出とスペクトル中の各調波の振幅 A_v, 位相 ϕ_v の計算を行う．

【時間領域リサンプリング】
時間領域のリサンプリングにより声質ターゲットサンプルの基本周波数 p_s を p_v と一致させる．このとき，声質ターゲットサンプルのリサンプリング係数 r は $r = p_v/p_s$ で表される．なお，この処理は，周波数領域においてはスペクトルの周波数方向の線形伸縮に相当する．音高補正をリサンプリング処理により行う最大の利点は，サブハーモニクスを含めた全てのスペクトル構造が一様に伸縮されるため，各調波とその周辺のサブハーモニクスの相対的な振幅・位相の関係の適切さが保たれる点にある．もし音高補正のために，従来の多くのボコーダのようにスペクトルが調波とノイズ成分のみから成るような仮定の下で処理を行うと，音高補正後にサブハーモニクスを適切に残すことが難しい．

【調波成分のマッピング】
スペクトルに含まれる全ての成分の一様な伸縮は，調波同士の間隔（音高）だけでなく調波成分の包絡やノイズ成分の包絡も伸縮させてしまうため，可能な限りリサンプリング処理前のスペクトルの概形構造を復元することが望ましい．そこで，伸縮したスペクトルの調波成分が，なるべく元サンプルの調波の位置に近くなるように再配置（マッピング）を行う．再配置後の調波のインデックス m_i は以下のように決定される．

$$m_i = \left\lfloor \frac{i}{r} + 0.5 \right\rfloor \qquad i = 0, 1, \ldots, N-1 \tag{1}$$

ここで N は調波の数を，i は調波のインデックスを表す．なお，ここでの処理は，phase-locked vocoder [17] の手法が適する．

【フィルタリング】

合成スペクトル Y の k 番目の周波数 bin の値はその最近傍の調波の index を i とすると以下のように定められる．

$$Y[k] = \widehat{S}[k + d_i] g_i e^{j\theta_i} \tag{2}$$

$$d_i = \left\lfloor p_v \left(m_i - i\right) \frac{L}{f_s} + 0.5 \right\rfloor \tag{3}$$

ただしここで $\widehat{S}$ はリサンプリング処理後の声質ターゲットサンプルのスペクトルを，d_i は i 番目の調波成分に関する単位を bin とする周波数シフト量を，L はフレームの分析窓長を，f_s はサンプリング周波数を表す．g_i および θ_i は i 番目の調波成分に対する振幅および位相の補正量であり，g_i がスペクトル包絡（音色）に関して，θ_i が位相に関して，それぞれ声質ターゲットサンプルを入力音声に近づけるためのファクターとして働く．その値は

$$g_i = \frac{A_v[i]}{A_s[m_i]} \tag{4}$$

$$\theta_i = \frac{\phi_v[i]}{\phi_s[m_i]} \tag{5}$$

として求められる．なお，A_v および A_s は各調波成分に相当する帯域における，振幅の重み付き平均により求められる．注意点として，フレームバイフレーム処理が時間方向に対して逆方向に進行している際は，計算される調波の位相を反転させなければならない．

【スペクトルモーフィング】

ここまでの処理により得られた「音高および調波成分の振幅と位相に関して，入力音声と同じものを付与された声質ターゲットサンプル」は，声質ターゲットサンプルの励振源成分と入力音声の音色および音高を持つ信号となる．結果，そのスペクトルと入力音声のスペクトルを任意の割合で足し合わせることが，入力音声に対し任意の分量の声質ターゲットサンプルの声質成分を付与することとなる．

3.2 音高補正に伴う不自然さへの対策

3.2.1 オーバーサンプリングによるエイリアシング対策

$r > 1$，すなわち声質ターゲットサンプルの音高を高くする方向にリサンプリング処理を行った場合，周波数領域上で伸張されたスペクトルがナイキスト周波数を上回る領域に侵入し，エイリアシングを発生させる．もしこの問題を回避するために信号にあらかじめローパスフィルタをかけてしまうと，リサンプリング処理前の信号に失われる成分が発生し，合成時に調波のマッピングが行えなくなってしまう場合がある．これを解決するためには，オーバーサンプリングによりナイキスト周波数を引き上げる手段が有効である．ただしこれはフーリエ変換の計算コストの増加とのトレードオフとなる．

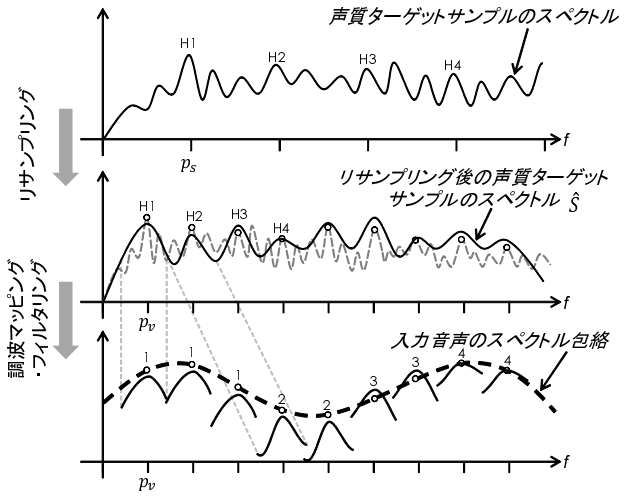

図 2 合成基本周期が分析窓長より長くなる場合の，リサンプリング処理及び調波マッピングの例

Fig. 2 An illustration of the process of resampling and harmonic mapping. .

3.2.2 マクロピリオドが分析窓長より長い場合の対策

分析窓長がマクロピリオドより十分長い場合はスペクトル中にサブハーモニクスが表れ，各調波は時間経過に対して比較的安定的な挙動を示す．しかしながら，音高を大きく下げる方向にリサンプリング処理を行い，マクロピリオドが分析窓長を超えてしまうと，サブハーモニクスは観測されなくなり，その代わり，時間経過とともにマクロピリオドに同期して調波のゲイン g_i に変調がかかるようになる．つまりこの場合，調波のゲインはフレームをまたいだ振幅揺らぎの影響を既にその中に含むものとなっているため，これを使用して毎フレームの振幅補正を行うと，合成音は結果的に時間方向の振幅変調が打ち消されたものになり，聴感上のグロウル系統の声質感が低下してしまう．この問題は声質ターゲットサンプルの分析窓長を十分長く設定すれば解決可能であるが，一方で，分析窓長を長くすることは，急峻な変化を持つ信号の表現力低下などの問題ももたらすため，合成システム全体の分析窓長を長くすればよいというものでもない．そこで本稿では，声質ターゲットサンプルの事前分析にのみ，長い時間長の分析窓を使用することを提案する[*4]．

この時の処理全体のイメージを図 2 に示す．あらかじめ十分長い分析窓長から得られたスペクトル（図 2 上段）は，リサンプリング処理により周波数方向に縮んでもサブハーモニクス構造を保ち（図 2 中段の破線），その調波（H1, H2, H3, ...）は時間経過に対して比較的安定的な挙動を示

[*4] 本研究では，BPM120, 4/4 拍子における全音符程度の長さのサンプルを使用した．

す．一方，リサンプリング処理後の信号を実際の合成処理で使用される分析窓長で分析して得られるスペクトル（図 2 中段の実線）では，サブハーモニクスが観測されなくなり，各調波に対して時間経過とともにゲイン方向の変調が加わるようになる．その結果，図 2 の中段では合成処理用分析窓によるスペクトル（実線）の各調波のゲインが，長時間窓によるスペクトル（破線）の各調波のゲインと少しのずれを持っている．ここで，式 (2) のスペクトル補正処理のための計算に長時間窓によるスペクトル（破線）を使用する．それにより調波ゲインの時間方向の変調の影響を受けない補正処理を行うことができ，合成音声中にグロウル系統の声質を再現することが可能となる．

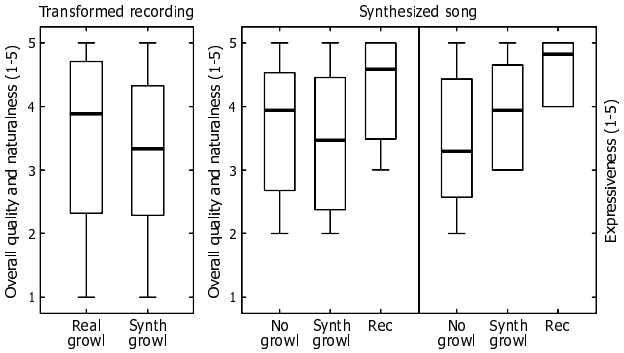

図 3 MOS による主観評価実験結果
Fig. 3 Results of the listening tests.

3.3 手法の特長

提案手法は本章冒頭の「満たすべき要件」に対し，以下のような解答を与える．

(1) 不規則な挙動が多くモデル化が難しいグロウル系統の音声に対して特別なモデルを仮定せず，サンプルを直接使用した合成を行うため，自然な合成結果が期待される
(2) 声質ターゲットサンプルからは音色の成分を取り除いた成分のみを使用するので，入力音声の発話内容そのものに影響を与えない．そのため，声質ターゲットサンプルは入力音声と同一の発話内容および時間同期を必要としない
(3) 「グロウル系統の励振源成分と入力音声の音色を持つ合成音声」と「入力音声」の間でスペクトルの振幅・位相合わせまで行っているため，単純にそれぞれのスペクトルを任意の比率で足し合わせることで，入力音声にグロウル系統の声質成分を任意の分量だけ付与することを実現する

4. 主観評価実験

グロウル系統の声質を持つ安定的に持続発声された音声のサンプルと，入力音声サンプルを使用して歌声合成および主観評価実験を行った．主観評価実験は 2 種類を行い，いずれも音楽・音声信号処理に携わる 17 名の被験者による 5 段階 Mean Opinion Score(MOS) 方式の評価による．

4.1 実験条件

実験 1 では，入力音声に自然歌唱を用いた．4 曲の自然収録からグロウルを含む 4 フレーズとグロウルを含まない 5 フレーズを抜き出し，含まないものに本手法によりグロウルを付与した．なお，1 フレーズは 2〜5 秒程度の長さを持ち，グロウルを付与する箇所と量は，歌声としてなるべく自然に聞こえるように人手で決定した．この，自然および合成のグロウルを持った 9 フレーズを評価用サンプルとした．使用した 4 曲の歌い手は男性 2 名，女性 2 名で，すべてのフレーズにおいて，声質ターゲットサンプルの提供者と入力音声の提供者が異なる（ただし性別は同一）．被験者は “全体の品質および自然性 (overall quality and naturalness)” に関して 1(miserable), 2(poor), 3(average), 4(good), 5(excellent) の 5 段階評価，および提示されたサンプルが自然音声か合成音声かの判断を行った．

実験 2 では，入力音声に市販の modal な声質を前提とした diphone ベースのサンプル接続型歌声合成ソフトウェア VOCALOID3 [18] の合成エンジンを使用して合成した歌声を使用した．聴感上なるべく自然な歌声を得るために，合成音声の音高および音素タイミング情報に関してはグロウルを含む自然歌唱の収録から抽出されたものを直接与えている．また，グロウルを付与する箇所及びモーフィング量については自然歌唱に極力似るように手動で決定した．被験者はブラインドで提示された，グロウルなし，グロウル付与，自然歌唱 の 3 種類のサンプルに対し，実験 1 と同様の 5 段階評価を行った．評価項目は “全体の品質および自然性 (overall quality and naturalness)” と “表現性 (expressiveness)” とした．

4.2 実験結果

主観評価実験の結果を図 3 に示す．太線が主観評価値の平均（MOS）を，箱の上端と下端が平均値の上側の標準偏差と下側の標準偏差を，箱の上下に伸びる線が最大値および最小値を表し，図の左側が実験 1 の，右側が実験 2 の結果を表している．実験 1 の結果では，合成によるグロウル音声の評価は，自然のグロウル音声の評価をやや下回るも，3 (average) よりやや高いところに位置した．表 1 に示される主観的識別実験の結果では，合成音声の提示に対して，半数に近い割合の回答が合成音声と確信を持てないことを示すという結果を得た．このことから，グロウルを再現する合成音声の自然性は，比較的良好であると考える．実験 2 の結果では，グロウルを本手法により付与しない場合とする場合において，自然性に関しては実験 1 と同様の傾向が見られた．一方で，表現性 (expressiveness) に関し

表 1 グロウルを含む自然音声と人工的にグロウルを付与された音声の主観的識別実験の混同行列

Table 1 Confusion matrix between recorded growl excerpts and excerpts with synthetic growl generated by morph.

		回答		
		自然音声	合成音声	わからない
提示	自然音声	66.17%	19.12%	14.71%
	合成音声	34.12%	54.12%	11.76%

てはグロウルを付与することにより上昇したことが確認された．このことから，本手法で付与したグロウルが，一定の自然性を持ちながら表現性の向上に寄与したことが確認できた．

5. まとめ

本稿では，グロウル系統の声質を持つ歌声を合成するための新しい手法を提案した．本手法では任意の入力音声から音色成分としてスペクトル包絡を，あらかじめ用意する声質ターゲットサンプルから声質に関する情報として励振源成分をそれぞれ用いて，音声の合成を行う．合成処理は声質ターゲットサンプルをベースとして，その音高と音色を入力音声を目標として合わせこむことで行われる．音高の合わせこみのために時間領域のリサンプリング処理を，音色の合わせこみのために各調波成分の振幅と位相に関して入力音声のそれを目標とした補正をそれぞれ行う．そうして得られたスペクトルと，入力音声のスペクトルを任意の割合で足し合わせることで，グロウル系統の声質を任意の分量だけ持った音声が合成される．本手法ではモデル化などは行わず，声質ターゲットサンプルを写実的に利用することで，高い自然性でターゲットの声質を再現している．また，パラレルレコーディングを必要としないこと，任意の箇所に任意の分量だけ声質を付与できることなどのメリットから，実用的な歌声合成システムへの応用が期待できる．

本稿ではターゲットとしてグロウル系統の声質のみに言及したが，本手法の大きな利点の一つとして，その他の声質にも広く応用可能であるという点が挙げられる．例えば，“tense”（ハリがある）や “breathy”（息成分が多い）といったような，信号の基本周波数の抽出に大きな問題が無いような声質であれば，直接この手法が使用できると考えられる．ただし実際には，歌声の静的な表現成分全体の再現には，声質の変化だけでなく音色の変化も伴わないと不十分である場合も多い．それらを適切に与えられるようにすることは今後の課題である．また，エクストリームメタルなどで使用されるいわゆるデスボイスと呼ばれるような，より信号の不規則性が大きく基本周波数の抽出が難しいような声質の再現も，課題の一つである．

参考文献

[1] Gobl, C. and Ní Chasaide, A.: The role of voice quality in communicating emotion, mood and attitude, *Speech communication*, Vol. 40, No. 1, pp. 189–212 (2003).
[2] 石井カルロス寿憲，石黒　浩，萩田紀博：韻律および声質を表現した音響特徴と対話音声におけるパラ言語情報の知覚との関連 (音声言語, ＜特集＞情報処理技術のフロンティア)，情報処理学会論文誌，Vol. 47, No. 6, pp. 1782–1792 (2006).
[3] Sakakibara, K., Fuks, L., Imagawa, H. and Tayama, N.: “Growl voice in pop and ethnic styles”, *Proceedings of the International Symposium on Musical Acoustics* (2004).
[4] 加藤圭造，伊藤彰則：“グロウル及びスクリーム歌唱の合成に向けた音響的特徴の分析”，情報処理学会研究報告，2012-SLP-90, No. 14, pp. 1–6 (2012).
[5] 西脇裕展，坂野秀樹，旭　健作：“スクリーム唱法における基本周波数とスペクトル変動の相関の調査”，電子情報通信学会音声研究会 (SP)，京都 (2013).
[6] Schoentgen, J.: Stochastic models of jitter, *J. Acoust. Soc. Am.*, Vol. 109, pp. 1631–1650 (2001).
[7] Ruinskiy, D. and Lavner, Y.: “Stochastic models of pitch jitter and amplitude shimmer for voice modification”, *Proc. IEEE 25th Convention of Electrical and Electronics Engineers in Israel (IEEEI)*, pp. 489–493 (2008).
[8] Verma, A. and Kumar, A.: “Introducing Roughness in Individuality Transformation through Jitter Modeling and Modification”, *Proc. Acoustics, Speech, and Signal Processing (ICASSP)* (2005).
[9] Loscos, A. and Bonada, J.: “Emulating Rough And Growl Voice In Spectral Domain”, *Proc. 7th Int. Conference on Digital Audio Effects (DAFX)* (2004).
[10] Cano, P., Loscos, A., Bonada, J., de Boer, M. and Serra, X.: “Voice Morphing System for Impersonating in Karaoke Applications”, *Proc. 2000 International Computer Music Conference (ICMC)* (2000).
[11] Morise, M., Onishi, M., Kawahara, H. and Katayose, H.: “v.morish’09: A Morphing-Based Singing Design Interface for Vocal Melodies”, *Proceedings of the 8th International Conference on Entertainment Computing*, Berlin, Heidelberg, Springer-Verlag, pp. 185–190 (2009).
[12] Yonezawa, T., Suzuki, N., Mase, K. and Kogure, K.: “Gradually Changing Expression of Singing Voice based on Morphing”, *Proc. Interspeech*, pp. 541–544 (2005).
[13] 河原英紀，森勢将雅：歌声を見て触る: TANDEM-STRAIGHT と時変モーフィングが提供する基盤，情報処理学会研究報告. [音楽情報科学]， Vol. 2010, No. 6, pp. 1–6 (2010).
[14] Roebel, A., Huber, S., Rodet, X. and Degottex, G.: “Analysis and modification of excitation source characteristics for singing voice synthesis”, *Proc. Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5381–5384 (2012).
[15] Lu, H.-L. and Smith, J. O.: “Glottal source modeling for singing voice synthesis”, *Proc. 2000 International Computer Music Conference (ICMC)*, pp. 90–97 (2000).
[16] Kawahara, H. and Morise, M.: “Analysis and synthesis of strong vocal expressions: Extension and application of audio texture features to singing voice”, *Proc. Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5389–5392 (2012).
[17] Laroche, J.: “Frequency-Domain Techniques For High-Quality Voice Modification”, *Proc. 6th Int. Conference on Digital Audio Effects (DAFX)* (2003).
[18] `http://www.vocaloid.com/`.