Vol.2012-MUS-94 No.14
Vol.2012-SLP-90 No.14
2012/2/4

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

加藤圭造[†1]　伊藤彰則[†1]

本研究ではデスメタル，メタルコアなどエクストリームメタルと言われるジャンルで頻繁に用いられる，グロウル及びスクリーム歌唱について音響的特徴の分析を行った．先行研究で特殊な発声の音響的な特徴として示されたサブハーモニクスの存在やmacro pulse 構造の調査，病的音声の分析になどに使われる jitter，shimmer，HNR の値について測定を行った．

Acoustic analysis towards extreme voice synthesis of death growl and scream singing voices

Keizo Kato[†1] and Akinori Ito[†1]

In this study, we analized acoustic feauture of growl and scream singing voices used in extream metal music, such as death metal, metal core, and so on. We observed sub-harmonics and macro pulse structures those are reported as accoustic features of rough voice. We also measured jitter, shimmer, and HNR values.

1. はじめに

近年，歌声合成に関する研究は盛んに行われており，商用歌声合成システムを利用した楽曲製作は盛んに行われている．製作される楽曲のジャンルは多岐に渡り，ポップス，ロックなどから，ハウス，エレクトロニカ，トランスなどの電子音楽，果てはデスメタル，メタルコア等，エクストリームメタルと呼ばれるジャンルまで様々である．

元来人間がそれぞれの楽曲で用いる歌唱表現や声質は多様である．例えばボサノバやシャンソンでは通常の音声よりウィスパーボイス (息漏れ声) が好んで用いられる傾向があり，またロックでは通常よりもしわがれた声質の所謂ハスキーボイスや，だみ声と言われる声が通常音声よりも好まれることがある．エクストリームメタルの分野ではしばしば，デスボイスと言われる声が楽曲の中の大部分に用いられる．

歌声合成システムを用いて製作された楽曲には，曲調に合わせて合成された歌声の声質を加工し，その分野の楽曲で用いられている唱法に近づけようと試みたものが多く存在する．特にエクストリームメタルでは，楽曲中の大部分がデスボイスで歌唱されるため，デスボイスは必須とも言える．その為デスボイスを模そうとするユーザーは多い．特殊な声質や唱法による歌声，特にデスボイスを合成することは，歌声合成システムを用いて楽曲製作を行う多くのユーザーから望まれていると考えられる．

1.1 デスボイス

デスボイスとは，デスメタルなど，エクストリームメタルの楽曲中の大部分で用いられる，非常に荒々しく，邪悪な印象を持つ声の日本における通称である．デスボイスは聴感上の違いにより，慣習的にいくつかの声種に分類される．声種の分類は個人やコミュニティにより一様ではない．本研究では以下のようにデスボイスを分類することにする．

- グロウル ･･･　聴感上低く，暗い印象を持つ．基本周波数をほとんど認識できない声と言われる．日本以外では単にグロウル (growl) ではなく，death growl，または grunt と呼ばれることも多い．CANNIBAL CORPS のボーカル Chris Banes などがこの声で歌唱をしている．
- スクリーム ...グロウルに比べ高い声として感じる声である．また，グロウルに比べて基本周波数を認識しやすい．ALL THAT REMAINS のボーカル Philip Labonte や，AUGUST BURNS RED のボーカル Jake Luhus 等がこの声を主に用いて歌唱をしている．
- ガテラル ...グロウルより更に低く，暗い印象を持ち，基本周波数も認識しづらい声である．言葉の明瞭度も低い下水道のゴボゴボとした音の様に形容されることがよくある．息を吸うことにより発声されることも多い．CEREBRAL BORE のボーカル Simone Som Plujimers 等が楽曲中で用いることがある．
- ピッグスクウィール ...豚の鳴き声のような，独特のビキビキとした響きを持つ声で

†1 東北大学大学院工学研究科
Tohoku University

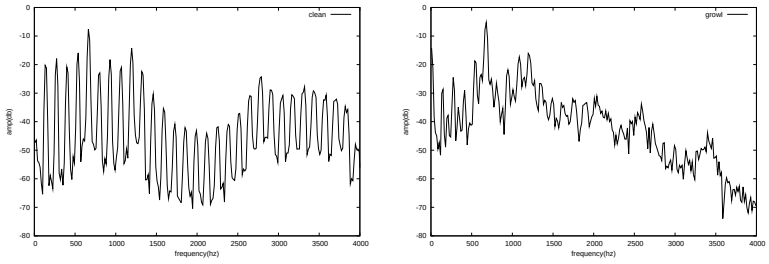
Vol.2012-MUS-94 No.14
Vol.2012-SLP-90 No.14
2012/2/4

ある．他の声種に比べ使用されることは少ない．JOB FOR A COWBOY のボーカル Jony Davy 等が楽曲の中で用いることがある．

本研究ではこれらのうち，グロウル，スクリームに対し，合成に向けた音響的特徴の分析を行う．また，グロウルは Louis Armstrong などのジャズシンガーが用いる，がなりのような，通常より荒々しい印象の唱法を呼ぶことがある．ゴスペルのシャウト唱法や，演歌の唸り唱法等が同様の唱法であると考えられるが，本研究ではこの唱法を区別のためポップグロウルと呼ぶことにする．

2. 先行研究

ポップグロウル音声などの特殊な発声においては，声帯ヒダ以外の器官も振動し，音源が作られることが分かっている[⋆1]．Sakakibara らは高速デジタル画像処理による分析と EGG(Electroglottography) 分析により，ポップグロウル音声の発声機構を調査した[1]．その結果，ポップグロウル発声時には声帯ヒダだけでなく，声帯ヒダ上部の披裂喉頭蓋ヒダが強く振動し，声帯ヒダの周期振動が持つ調波成分にサブハーモニクスが加わり，音源特性が変化することが明らかになった．図 1(a) は通常音声/a/のスペクトル，図 1(b) はポップグロウル音声/a/のスペクトルである．



(a) 通常音声/a/のスペクトル　(b) ポップグロウル音声/a/のスペクトル

図 1 通常音声とポップグロウル音声のスペクトル

⋆1 通常，人間の音声は呼気が声帯を通るときに，声帯ヒダが振動することにより生じる声帯音源波が，喉から唇までの空間 (声道) を通ることにより周波数特性が特徴づけられ，音色が変化し形成される．

また，Nieto はロックやメタルなどで使われる特殊な発声 (Extream Voice Effects:以下 EVE) の音響的特徴を分析することにより，通常音声を EVE 音声に変換する研究を行った[2]．Nieto の研究において EVE は Distortion，Rattle，Growl，Grunt，Scream に分類され，Growl は本研究におけるポップグロウルに，Grunt は本研究におけるグロウルに，Scream は本研究におけるスクリームに相当する．Nieto は音声波形を観察することにより，EVE 音声は大きなパルス (macro pulse) の中に複数の小さなパルス (small pulse) が含まれる，macro pulse 構造を持つことを発見した．例えば，ポップグロウルは macro pulse の中に平均して 3 つの glottal pulse を含む．ポップグロウルの母音/e/の音声波形を図 2 に示す．また，グロウル音声は macro pulse の中に 3〜6 個の glottal pulse がランダムに含まれる．スクリーム音声についての分析はされていない． Nieto は macro pulse 内に含まれる glottal

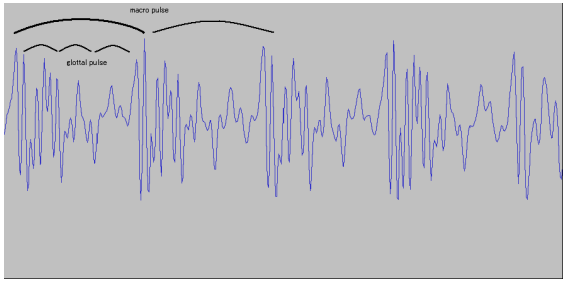


図 2 ポップグロウルのマクロパルス構造

pulse それぞれの周波数特性を分析し，得られた周波数特性を再現するフィルターを通常音声の音声パルスに対してかけることにより，通常音声から EVE 音声への変換を試みた．

本研究ではこれらの音響的特徴に加え，グロウル及びスクリーム音声の合成にあたり有用と思われる特徴について調査を行った．

3. 音響的特徴の分析

3.1 音声収録

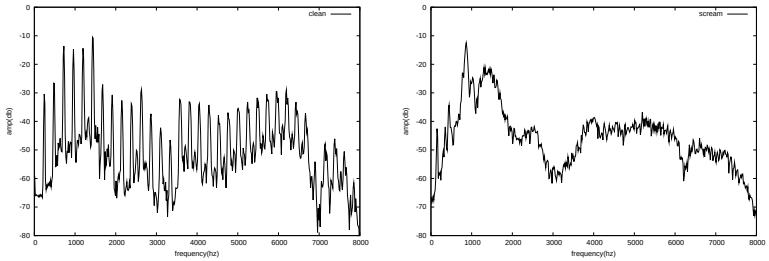
グロウル音声，スクリーム音声の音響的特徴を分析するため，音声収録を行った．音声の収録はグロウル音声，スクリーム音声のそれぞれについて 1 名ずつ行った．音声は/a/,/i/,/u/,/e/,/o/母音を単音で発声したものを 3 回ずつ収録した．また，合わせて通常

Vol.2012-MUS-94 No.14
Vol.2012-SLP-90 No.14
2012/2/4

音声についても収録を行った．それぞれの音声収録の際，歌唱者にとって出しやすいような発声で，なるべく一定な発声を心がけるように指示を行った．収録は防音室内で行い，オーディオインターフェースは EDIROL の UA-101 を，マイクは SHURE の SM58 を用いた．サンプリング周波数は 16kHz，量子化制度は 16bit であった．

3.2 スペクトル分析

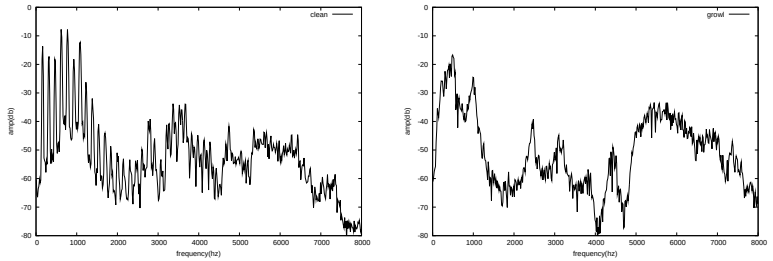
図 3(b) にスクリーム音声/a/の，図 3(a) に同歌唱者の通常音声/a/の周波数振幅スペクトルを，図 4(b) にグロウル音声/a/の，図 4(a) に同歌唱者の通常音声/a/の周波数振幅スペクトルを示す．FFT 点数 1024 点，フレームシフト 10ms，分析窓 Hamming 窓で 10 フレームに対してフーリエ変換を行った平均である．



(a) 通常音声/a/のスペクトル　(b) スクリーム音声/a/のスペクトル

図 3 スクリーム音声と通常音声のスペクトル

どちらの歌唱者に関しても通常音声は基本周波数の整数倍の周波数毎にピークが生じる調波構造持つのに対し，スクリーム音声は 135〜150Hz 付近とその約 3 倍の 400〜450Hz，また非整数倍の 350Hz 付近にピークが見られた．ポップグロウル同様，サブハーモニクスが加わっており，130〜150Hz の 2 倍である周波数域のピークはサブハーモニクスに埋もれていることが考えられる．図 1 のポップグロウルのスペクトルでは 1000〜1500Hz あたりの帯域にも周期的なピークが見られたが，スクリーム音声に関しては 1000Hz 以上の帯域でほとんどノイズ的であった．グロウル音声は全帯域でほとんどノイズ的であるといえる．いずれの母音に関しても，おおよそ同じような傾向が見られた．また，母音ごとのスペクトル包絡の変化を調べるため，ケプストラム距離を計算した．ケプストラム距離 CD は以下の式で表せる．



(a) 通常音声/a/のスペクトル　(b) グロウル音声/a/のスペクトル

図 4 通常音声とグロウル音声のスペクトル

$$CD = \frac{10}{\ln 10}\frac{1}{T}\sum_{t=0}^{T}\sqrt{2\sum_{i=1}^{M}(c[i]-c'[i])/M} \quad (1)$$

ここで $c[i]$，$c'[i]$ は音声のケプストラム，M はケプストラム次数，T はフレーム数である．フレーム幅 256 点，フレームシフト 5ms，Hamming 窓，ケプストラム次数 24 次で計算したスクリーム音声と同歌唱者の通常音声，グロウル音声と同歌唱者の通常音声の各母音のケプストラム距離を表 1 に示す．

表 1 ケプストラム距離

	a	i	u	e	o
scream-clean	6.097	7.066	6.815	6.661	6.207
growl-clean	11.154	11.645	10.273	10.391	11.335

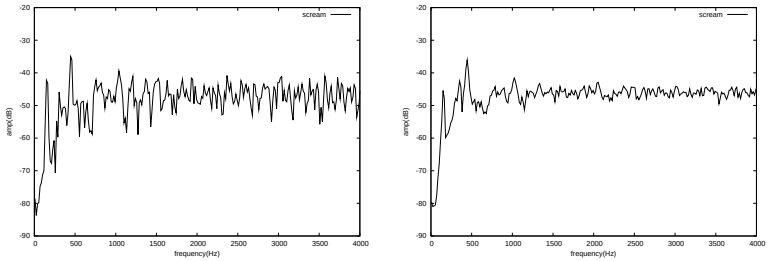
ケプストラム距離の点からスクリーム音声よりもグロウル音声の方が通常音声に対してのスペクトル包絡の変化が大きいといえる．また，スクリーム音声は/a/，/o/母音で他の母音よりもケプストラム距離が小さく，グロウル音声は/u/，/e/母音で他の母音よりもケプストラム距離が小さくなった．

3.3 音源特性

声道フィルタの影響を除去し，より明確に音源の特性を調べる為に，収録した音声信号に対し線形予測分析[3] を用い，残差信号を抽出した．予測次数は 25 次である．予測残差として抽出された音源信号のスペクトルの例を図 5，図 6 に示す．FFT 点数 1024 点，フレーム

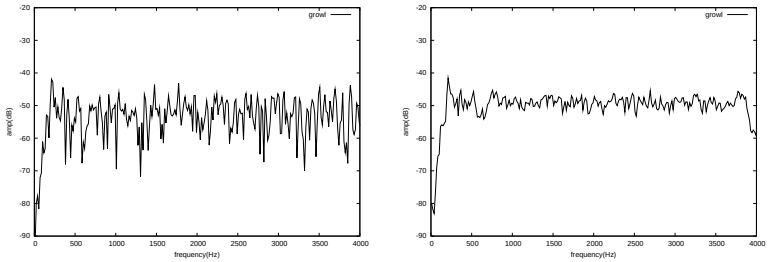
Vol.2012-MUS-94 No.14
Vol.2012-SLP-90 No.14
2012/2/4

シフト 10ms，短時間平均スペクトルは分析窓 Hamming 窓で 10 フレームの振幅スペクトルの平均，長時間平均スペクトルは 70 フレームの平均である．



(a) 短時間平均スペクトル　　(b) 長時間平均スペクトル

図 5　スクリーム音声/a/の残差信号のスペクトル



(a) 短時間平均スペクトル　　(b) 長時間平均スペクトル

図 6　グロウル音声/a/の残差信号のスペクトル

図 5 において，スクリーム音声は 442Hz にもっとも大きなピークが見られ，そのおよそ 3 分の 1 の 147Hz，整数倍の関係に無い 355Hz にピークが見られた．442Hz の整数倍にピークのようなものは見られなかった．また，聴感上 442Hz 付近をピッチとして知覚することができた．図 6 において，グロウル音声の残差信号の短時間平均スペクトルはほとんどノイズ的ではあるが，ホワイトノイズとは異なり，200Hz 以下の振幅が小さくなっている．長時間平均スペクトルを見るとわずかながらではあるが，200Hz 付近の振幅が他の周波数域よりも大きくなっている．聴感上も 200Hz 付近をピッチとして知覚することができた．従来グロウル音声はほとんどピッチが無い声や非常に低いピッチを持つ声とイメージされることが多かったが，実際には成人男性が通常会話で用いる音域とあまり変わらない 200Hz 付近にピッチらしきものがあることが分かった．

3.3.1　macro pulse 構造

2 章で述べたとおり，EVE 音声は複数の大きなパルス (macro pulse) の中に小さなパルス (glottal pulse) が複数個含まれる macro pulse 構造を持ち，グロウル音声は macro pulse の中に glottal pulse が 3〜6 個ランダムに含まれていると言われている．この様なマクロパルス構造が収録した音声内に含まれるか調査を行った．図 7 にスクリーム音声/a/の残差信号波形を，図 8 にグロウル音声/a/の残差信号波形を示す．高周波域におけるノイズの影響を除去するために，ローパスフィルターにより平滑化を行っている．カットオフ周波数は 1000Hz である．

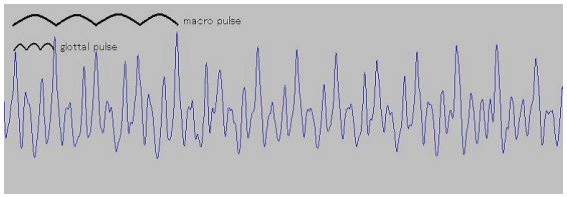


図 7　スクリーム音声/a/の残差信号波形

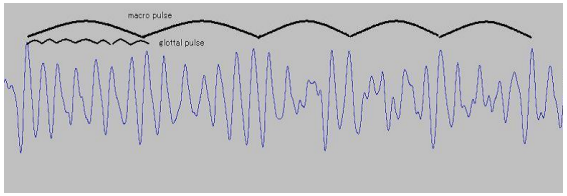


図 8　グロウル音声/a/の残差信号波形

スクリーム音声に関して，多くの区間で 3 つの glottal pulse が macro pulse に含まれる構造が観測できた．グロウル音声に関してもおおよそ Nieto が述べたような構造が観測できた．

3.3.2 jitter，shimmer 測定

jitter,shimmer はそれぞれ基本周期，振幅の揺らぎを表す指標である．病的音声の診断などに用いられ，jitter や shimmer が大きくなると粗造性 (Rough) が大きくなることが知られている[5]．jitter，shimmer は以下の式で定義される[4]．

$$jitter(absolute) = \frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i-1}| \tag{2}$$

$$jitter(relative) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i-1}|}{\frac{1}{N}\sum_{i=1}^{N}T_i} \tag{3}$$

ここで T_i は i 番目の音源基本パルス間隔，N はパルス間隔の数である．

$$shimmer(dB) = \frac{1}{N-1}\sum_{i=1}^{N-1}|20\log(A_{i+1}/Ai)| \tag{4}$$

$$shimmer(relative) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|A_i - A_{i-1}|}{\frac{1}{N}\sum_{i=1}^{N}A_i} \tag{5}$$

ここで A_i は抽出された振幅のピーク，N は抽出されたピークの数である．

また，病的音声の分析などにおいて，周期と，隣接する周期 3 点の平均の揺らぎである RAP，周期と隣接する周期 5 点の平均である PPQ，振幅と，隣接する振幅 11 点の平均の揺らぎである APQ がしばしば用いられる．これらの指標について分析を行った．RAP は以下の式で表され，同様にして PPQ，APQ が求められる．

$$RAP = \frac{\frac{1}{N-2}\sum_{i=2}^{N-1}|(T_{i-1}+T_i+T_{i+1})/3 - T_i|}{\frac{1}{N}\sum_{i=1}^{N}T_i} \tag{6}$$

基本周期は，全てのゼロクロス点間隔の平均と長時間スペクトルのピークから，スクリーム音声に対しては 350Hz〜500Hz，グロウル音声は 150〜250Hz であると考え，その範囲で検出を行った．振幅のピークは抽出された基本周期内での最大振幅とした．スクリーム及びグロウル音声の残差波形はノイズの影響が大きく，基本周期の抽出が困難であったため，ローパスフィルターにより平滑化を行ってから測定を行った．それぞれ音声/a/の残差波形から，中間部の発声が安定していると思われる 100 周期を用いた．また，スクリーム音声に関しては macro pulse に対しても測定を行った．用いた周期は 40 周期である．測定した結果を表 2 に示す．グロウル，スクリーム音声ともに，同歌唱者の通常音声の各指標を大きく

表 2 shimmer,jitter 測定値

	scream(g)	scream(m)	growl	clean(s)	clean(g)
jitter(absolute)(msec)	0.326	0.206	1.283	0.856e-2	0.283e-3
jitter(relative)(%)	13.919	3.011	27.817	0.204	0.438
RAP(%)	7.602	1.764	15.517	0.128	0.230
PPQ(%)	9.907	1.570	18.116	0.105	0.247
shimmer(absolute)(dB)	5.563	2.161	4.295	0.314	0.509
shimmer(relative)(%)	56.768	22.949	41.840	3.613	5.844
APQ(%)	38.761	15.297	25.033	2.545	4.088

scream(g)：スクリームの glottal pulse :scream(m)：スクリームの macropulse
clean(s)：スクリーム歌唱者の通常音声 :clean(g)：グロウル歌唱者の通常音声

上回った．growl 音声については基本周期の抽出をヒューリスティックに行った為，基本周期の揺らぎとして正しく測定できていない可能性が高い．スクリーム音声に関して，macro pulse で測定を行っても同様のことが言える．病的音声の研究において，PPQ の正常値は 0.13〜1.00%，APQ の正常値は 0.75〜3.37%と言われており，この値も大きく上回っている．グロウル話者の通常音声について，APQ が正常値を若干上回っているが，残差波形を用いたことによる影響が現れている可能性がある．

3.3.3 HNR 測定

HNR(harmonics-to-noise-ratio) は周期信号における調波成分と非調波成分 (雑音成分) のエネルギーの比である．ポップグロウル音声では，雑音成分が大きくなり，HNR が小さくなることが知られている[7]．HNR は以下の手順で求められる．

音声信号 $x(t)$ に対し，遅延時間 τ における自己相関関数 $r_x(\tau)$ は以下のように定義される．

$$r_x(\tau) = \sum_{t=0}^{N-\tau} x(t)x(t+\tau) \tag{7}$$

τ を時間 l〜L とした時，自己相関関数 $r_x(\tau)$ が最大になる τ を τ_{max} とすると，時間 l〜L

Vol.2012-MUS-94 No.14
Vol.2012-SLP-90 No.14
2012/2/4

における最大正規化自己相関関数 $r'_x(\tau_{max})$ は

$$r'_x(\tau_{max}) = \frac{r_x(\tau_{max})}{r_x(0)} \tag{8}$$

である．ここで音声信号 $x(t)$ が調波成分 $H(t)$ と雑音成分 $N(t)$ の和からなっており，$H(t)$ と $N(t)$ の間に相関が無いとすると，音声信号 $x(t)$ の自己相関関数は調波成分 $H(t)$ と雑音成分 $N(t)$ の自己相関関数の和であると考えられる．分析区間における最大正規化自己相関関数 $r'_x(\tau_{max})$ は，信号内における調波成分の相対パワーであると考えられるため，調波成分と非調波成分はそれぞれ

$$H(t) = r'_x(\tau_{max}) \tag{9}$$

$$N(t) = 1 - r'_x(\tau_{max}) \tag{10}$$

と表すことができ，HNR は

$$HNR(\mathrm{dB}) = 10\log_{10}\frac{r'_x(\tau_{max})}{1 - r'_x(\tau_{max})} \tag{11}$$

と定義される．

スクリーム音声/a/と同歌唱者による通常音声/a/の残差波形，グロウル音声/a/と同歌唱者/a/による通常音声の残差波形，ホワイトノイズのそれぞれについて，HNR を測定した．分析窓幅 128 点，フレームシフト 64 点とし，フレーム毎に算出された HNR の平均を取っている．また，通常有声区間と無声区間を区別するため，最大正規化自己相関関数が閾値以下であった区間は HNR の算出に使わないが，今回の測定では閾値を設けずに行った．測定結果について，表 3 に示す．

表 3 HNR 測定値

	HNR(db)
scream	-4.992529
growl	-7.53103
clean(s)	9.114614
clean(g)	8.213341
whitenoise	-10.431067

clean(s) : スクリーム歌唱者の通常音声
clean(g) : グロウル歌唱者の通常音声

スクリーム音声よりもグロウル音声の HNR の値が小さくなっている．残差信号のスペクトルや，グロウル音声のピッチがスクリーム音声のピッチに比べて知覚しづらいことからも，この結果は妥当と思われる．

4. まとめと今後の課題

本研究ではグロウル及びスクリーム歌唱音声の音響的分析を行った．グロウル音声については先行研究で報告されていた macro pulse 構造の確認，スクリーム音声に関してはサブハーモニクスの確認に加え，先行研究では調査がされていなかった macro pulse 構造が発見された．また，病的音声の分析などに用いられる jitter，shimmer や HNR の測定を行った．その結果，通常音声よりも jitter,shimmer の割合が遥かに大きくなり，HNR の値は通常音声や先行研究で報告されているポップグロウル音声よりも小さい値になることが分かった．今後 jitter，shimmer について，それぞれの時系列の変動の分析や，より詳細な統計的分析をしていくことで，合成に有用な特徴が得られると考えられる．また，jitter，shimmer 計測の際，基本周波数の抽出が重要になる．グロウルに音声に関してはランダム性が大きく，基本周波数を一様に定める事が困難であると考えられる．基本周波数の決定に関して，手法の検討が必要である．また，今回分析のために収録した音声はそれぞれの発声について 1 人ずつのものしかなく，歌唱者の個人性に依存する分があると考えられる．歌唱者を増やすことにより，それぞれの発声についてより一般的な特徴を得ることが重要であると思われる．

参 考 文 献

1) K. Sakakibara, L. Fuks, H. Imagawa, N. Tayama.: Growl Voice in Ethnic and Pop Styles , Proc. Int. Symp. on Musical Acoustics, 2004.
2) O. NIETO : Voice Tranceformations for Extream Voice Effects, Master's thesis, POMPEU FABRA UNIVERSITY (2008).
3) 古井貞熙 : デジタル音声信号処理, 東海大学出版会 (1985).
4) M. Farrus, J. Hernando, P. Ejarque.: Jitter and Shimmer Measurements for Speaker Recognition, Proc. Interspeech, pp. 778-781(2007)
5) 兵頭 政光 : 音声障害の診断と治療 , 日本耳鼻咽喉科学会会報, Vol. 113, No. 10 pp.818-821(2010).
6) P. Boersma.: Accurate Short-term Analysis of the Fundamental Frequency and the Harmonics-to-noise Ratio of a Sampled Sound, Proc. Inst. Phonetic Science, vol. 17, pp. 97-110 (1993).
7) C. Tsai, L. Wang, S. Wang, Y. Shau, AND T. Hsiao.: Aggressiveness of the Growl-like Timbre: Acoustic Characteristics, Musical Implications, and Biomechanical Mechanisms, Music Perception, Vol. 27, No. 3 (2010)