Vol.2013-MUS-99 No.36
2013/5/12

歌唱合成音声のハスキーボイス化

佐々木　星太朗[†1]　　白木　善尚[†1]

歌唱合成音声とは，音声合成・デスクトップミュージック（DTM）ソフトウェア等を用いて，コンピュータ上で生成された歌唱音声である．本研究では，歌唱合成音声のハスキーボイス化の方法を提案する．ハスキーボイス化の基本的な手法は，音声信号の声帯情報である残差信号に処理を加える「残差変換」である．この「残差変換」について本研究では二種類の方法を提案している．さらに「残差変換」処理後の音質向上を図るためイコライザーの応用を試みる．また残差信号の振幅情報から「残差変換」処理のレヴェルを決定することによって，ハスキーボイスの自動生成を行なう．提案した二種類の「残差変換」法，イコライザー応用法，自動生成法，の４つの手法を用いて合成した音声の聴取実験を行い，魅力的なハスキーボイス化のためにはどのような手法や指標に有用性があるのかについて，検証する．

A method for transforming normal singing voice into husky synthesized singing voice

Seitarou Sasaki[†1]　and　Yoshinao Shiraki[†1]

The singing synthesized voice , is a singing voice using speech synthesis and the desktop music (DTM) software, etc., is generated on a computer. In this paper, we propose a method for transforming normal singing voice into husky synthesized singing voice. The residual signal is a vocal information of the speech signal. "This conversion to the residual signal" is the basic techniques of husky voice conversion. In this study, we propose two methods for this "residual conversion". In order to improve sound quality of the "residual conversion" treatment, we try to apply the equalizer. Further by determining the Revel "residual conversion" process from the amplitude information of the residual signal, we perform the automatic generation of husky voice. For generating synthesized speech using the four methods of the two types of proposed "residual conversion" method, the equalizer application method, an automatic generation method, we do listening tests. We do verification for two things; Which approach or increase the degree of attraction husky voice ? What indicators suitable to measure the degree of glamor husky voice ?

1. はじめに

歌唱合成音声とは，音声合成・デスクトップミュージック（DTM）ソフトウェア等を使用し，コンピュータ上で生成された歌唱音声である．代表的な例として，クリプトン・フューチャー・メディアから発売された「初音ミク」や飴屋／菖蒲（あめや・あやめ）制作によるフリーウェア「UTAU」が挙げられる[1]．近年，デスクトップミュージック（DTM）やエレクトロダンスミュージック（EDM）の普及・流行に伴い，これらの歌唱合成音声が使用される機会が増え，認知度も高まっている．そして，これらの音声を肉声に近い歌声として使用したいという需要が高まる事も予想される．

本研究では肉声に近い歌唱合成音声の生成を目標に置く．目標達成のための一つの側面として，独特な歌声，歌い方などと言った楽譜上に表せない要素，いわば「人間的魅力」を歌唱合成音声に付加する事が挙げられる[2]．この人間的魅力を持つ声の一つとして挙げられるのがハスキーボイスである．ハスキーボイスは，オペラなどで耳にする美しい歌声とは異なり，「しゃがれ」や「かすれ」といった一見，美しさと相反するような性質を持つ．しかし，哀愁などいった人間的魅力を感じさせる場合があり，実際にハスキーボイスを持つ多くの有名なアーティストが存在する．本研究では，特に，このハスキーボイスに焦点を当て，歌唱合成音声をハスキーボイス化する方法を提案する．また，ハスキーボイス化を通し，人間的魅力を指標化・特徴化する事も本研究の目的である．

本研究では，ハスキーボイスを次のように捉える．ハスキーボイスとは，かすれた声，ダミ声であり，主に声帯から生じる雑音を含む声である．声帯の器質的病変や運動麻痺から生じる症状である嗄声が，かすれた声，ダミ声などの，音色に関する総合的な声の異常を指す為[3]，ハスキーボイスも嗄声の一形態と見なせる．類似した声にデスボイスなどが挙げられるが，これは声帯とは別の器官から生じた音が含まれる声である為，ハスキーボイスとは区別されることが多い[4],[5]．また図 1 の例に示すように，ハスキーボイスの音声[a]はハスキーボイスではない音声[b]と比べ，周波数成分が広帯域に分散分布している事が確認される．

†1 東邦大学
Toho University

[a]ハスキーボイスの代表的な例として，ｻｻﾞﾝｵｰﾙｽﾀｰｽﾞの桑田佳祐氏の歌声の一部を使用している．
[b]ハスキーボイスではない音声として第一著者自身の歌声を使用している．

ⓒ 2013 Information Processing Society of Japan

Vol.2013-MUS-99 No.36
2013/5/12

本研究では，この周波数成分の広帯域分散分布をハスキーボイスの主要な特徴と見なす．

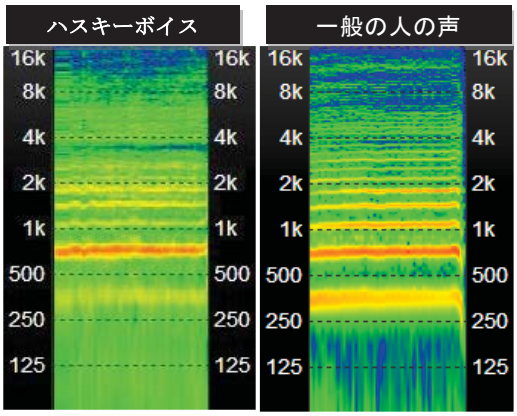

図 1　周波数成分分析によるハスキーボイスの比較図[c]
Fig. 1 Comparison with a normal voice and husky voice with a frequency component analysis.

本研究では，歌唱合成音声を線形予測分析（LPC）する事によって得られる残差信号に注目し，残差信号に変換，処理を加える事によってハスキーボイス化を行う．また，音質面，汎用性を考慮した方法の考案も行う．音質の保持にはイコライザーの技術を応用し，汎用性の面では，ハスキーボイスの自動生成を試みる．自動生成の技術は，土師知行,楯敬蔵,片岡英幸,藤木暢也著「残差波形分析による嗄声の客観的評価」文献[6]を参考としている．文献[6]で記述された「嗄声の度合いを測る指標として，患者の残差波形のヒストグラムを使用する」という方法をヒントとして，ハスキーボイス化における魅力度を計る指標，自動生成の手法を考案する．最後に，提案した二種類の「残差変換」法，イコライザー応用法，自動生成法，の４つの手法を用いて合成した音声で聴取実験を行い，その有用性に関して実証を試みる．

2. ハスキーボイス化の基本的手法

(1) 基本的手法

歌唱音声を線形予測分析（LPC）すると，線形予測係数と残差信号が得られる[7]．入力信号がこのような歌唱音声の場合，線形予測係数は歌詞部分・声道情報であり，残差信号はメロディ部分・声帯情報であると見なせる．前項で記述したように，ハスキーボイスは声帯から生じる雑音を含む音声である為，本研究では声帯情報である残差信号に処理・変換を加えた後，元の音声の線形予測係数と再合成する「残差変換法」をハスキーボイス化の基本的手法としている．次の図 2 に「残差変換法」を基本とするハスキーボイス化の処理フローを示す．

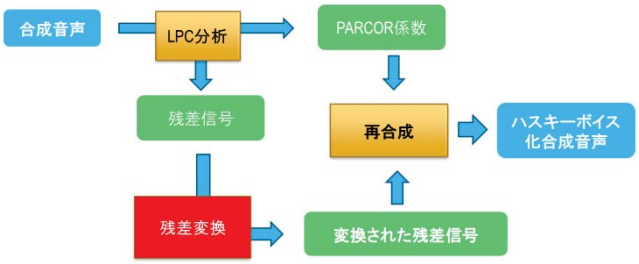

図 2　残差変換法に基づくハスキーボイス化の処理フロー
Fig. 2 Processing flow of the husky voice of based on residual conversion method.

本研究では，この残差変換法に関して以下に述べる二つの手法を検討する．

(2) 残差変換法１（ノイズ強調法）

声帯情報である残差信号は，周期性のある振幅の大きいパルス成分と，それ以外の振幅の小さいノイズ成分に大別できる（図 3）．ノイズ強調法は，このノイズ成分の振幅を増幅させる事により，再合成後の音声のノイズを強調させ，ハスキーボイス化を図る手法である．

パルス成分とノイズ成分の判定には，残差信号内の振幅の絶対値を使用する．振幅の値が絶対値の平均以上の場合，パルス成分と判定，平均以下の場合，ノイズ成分と判定する（図 4）．

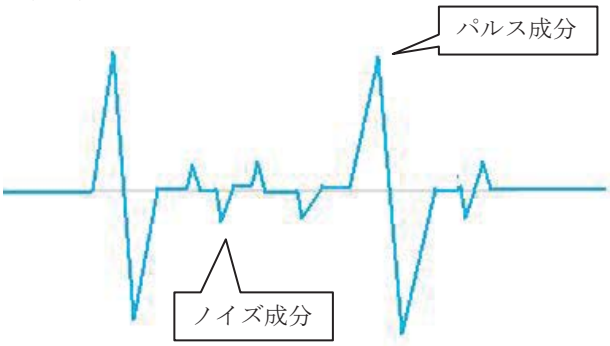

図 3　パルス成分とノイズ成分のイメージ図
Fig. 3　Image of pulse and noise components.

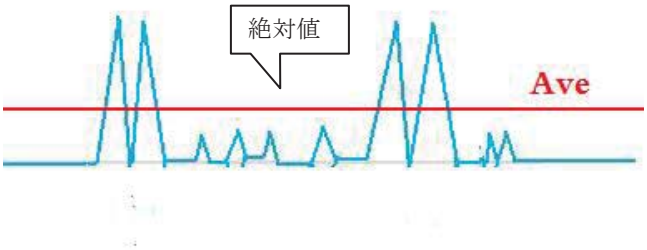

図 4　パルス成分とノイズ成分の判定のイメージ図
Fig. 4　Image of pulse and noise components at the time of judgment.

[c]周波数解析にはフリーソフト「Sound Engine」を使用した

ⓒ 2013 Information Processing Society of Japan

(3) **残差変換法２（間引き法）**

声帯情報である残差信号は，コンピュータ上においてサンプリング周波数[d]で表現されたデジタルデータである．このサンプルの中からランダムにサンプルを抽出し，データ値を 0 値に置き換える（図 5），すなわち間引きする事によって残差信号の周期性を一部無くし，疑似的ノイズを生じさせる手法がこの間引き法である．間引き処理のレヴェル調整は，全体のデータ数の何%を 0 値に置き換える（間引きする）かを手動入力により指定する．

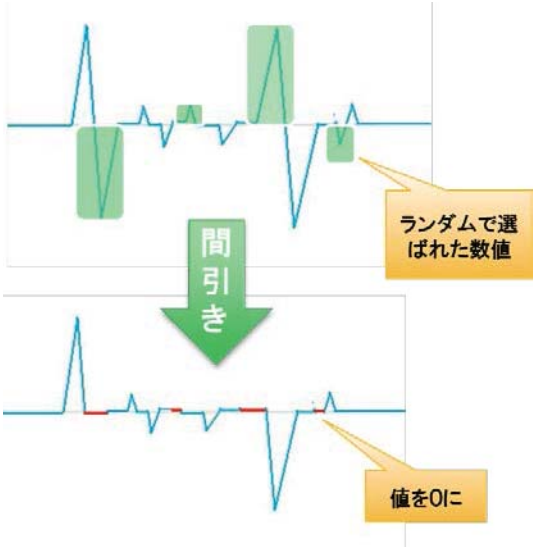

図 5　間引き処理のイメージ図
Fig. 5　Image of thinning processing.

3. 残差変換の応用法

3.1 イコライザー技術の応用

ハスキーボイス化において最も危惧しなければならない要素は音質の劣化である．ハスキーボイスは元来，雑音を含む音声である為，ハスキーボイス生成を試みるとどうしても，余計な雑音の成分が増え，音質の劣化に繋がってしまう．したがって，ハスキーボイス化の際には音質保持が最優先事項となる．

この音質保持については，残差変換時にイコライザーの技術を応用する事によって一定の効果が確認できた．これは一定の周波数のみに残差変換を行い，原音の下地を一部残すことによって，音質を保持するという手法である（図 6）．以後，この手法を「イコライザー応用法」と呼ぶ．

イコライザー応用法では，ハスキーボイス化に用いる原音を FIR フィルタによって設計されたバンドパス及びストップフィルタを用いて，残差変換を行う帯域と，それ以外の帯域に分離する．分離した一方の音声に残差変換・再合成を行った後，この二つを足し合わせる事により，一定周波数のみに残差変換を行うことが可能となる（図 7）．尚，現状では倍音付近の周波数帯域にハスキーボイス化処理を行い，其音と高周波数域の雑音部分を元音のまま保持する方法が最適と判断している．

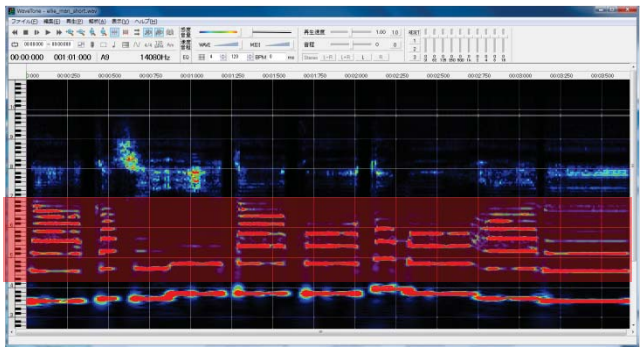

図 6　分離する周波数帯域を示した図[e]
Fig. 6　Diagram showing a frequency band to separate.

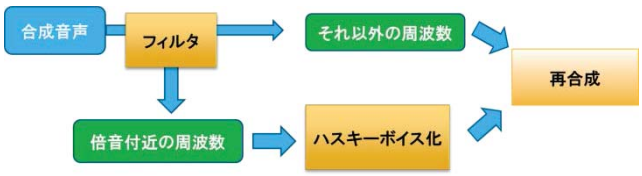

図 7　イコライザー応用法の処理フロー
Fig. 7　Processing flow of the application equalizer method.

3.2 ハスキーボイスの自動生成

間引き法の応用として，間引き処理のレヴェルを自動的に決定する事により，ハスキーボイスを自動生成する「自動生成法」を考案した．これを次の（1）（2）（3）に記述する．

(1) **間引きレヴェル決定指標**

自動生成の際のレヴェル決定指標には残差信号の振幅のヒストグラムを利用する．これは一般の人の声と比べ，ハスキーボイスの残差信号の振幅の分布が大きく異なる為である．本研究では桑田佳祐氏の声を理想的なハスキーボイスのモデルとし，振幅のヒストグラムにおいてモデルとの誤差を算出，この誤差を指標として最適な処理のレヴェルを決定している．また，声においての残差信号の振幅の分布が Rayleigh 分布[8]に従うと仮定する事によって，パラメーター1 つ（以降，誤差 α）で誤差を表現する事が可能となった（図 8）．実際にヒストグラムと誤差を算出した際の結果が図 9 である．

[d]サンプリング周波数は 44100Hz

[e]周波数解析にはフリーソフト「Wave tone」を使用した

ⓒ 2013 Information Processing Society of Japan

Rayleigh 分布とは確率密度関数

$$p(x)=\lambda^2 x e^{-\frac{\lambda^2 x^2}{2}} \qquad x \geqq 0$$
$$=0 \qquad x<0$$

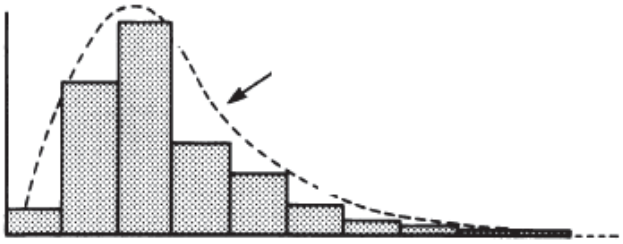

図 8　残差信号のヒストグラムと Rayleigh 分布
Fig. 8 Rayleigh distribution and the histogram of the residual signals.

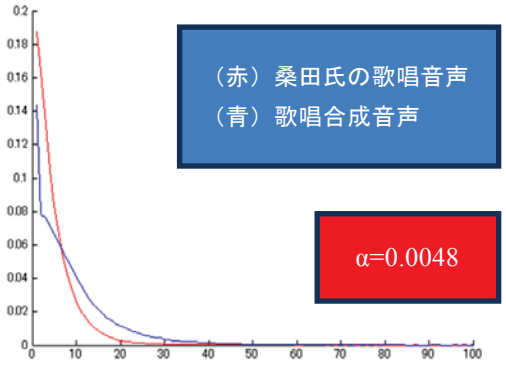

図 9　残差信号のヒストグラムと誤差算出
Fig. 9　Histograms of the residual signal and error calculation.

(2)　第一最小誤差による最適処理レヴェルの決定

(1)で記述した振幅のヒストグラムから算出する，誤差 α は，間引き処理のレヴェルを上げていくと，ある一定の数値までは下がり，それ以降は不規則なふるまいをする特性がある（図 10）．この誤差が最小になる初回のポイントを「第一最小誤差」と呼び，この第一最小誤差で処理レヴェルを決定する事により，振幅のヒストグラム上ではモデルと同じハスキーボイスの生成が可能となる．尚，第一最小誤差の算出は，検証の結果，間引きレヴェルを 1%ずつ上げるたびに判定を行う方法が最適であると，判明した．

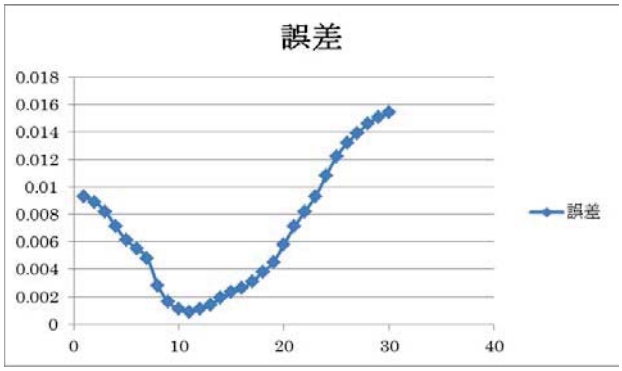

図 10　間引き度を 1%ずつ落としていった際の誤差の変化
Fig. 10　Error change when went down by 1% the degree of thinning.

(3)　歌唱区間割り出しによる処理制度の向上

歌唱を収録した音声データの場合，息継ぎの際の極めて振幅の小さい部分や，声を出していない無音の部分が必ず存在する．ハスキーボイス化を行う際，この無音部分にまで処理を施すと，予期せぬ雑音が付加されてしまうなどといった精度劣化の原因となる．また，第一最小誤差割り出しの制度も落ちてしまう．そこで，自動生成法の際には，振幅の小さい区間や無音の区間と，歌唱の区間を分離し，歌唱の区間のみに処理を行う事によって精度の向上を図る．

歌唱区間は，入力音声の振幅により算出する．任意の区間（現状では一秒間を 100 分割）の振幅の絶対値の平均を求め，この平均値が小さい部分を無音区間と判断し，歌唱区間の開始・終了点とする（図 11）．

こうして算出した各歌唱区間ごとに処理を行い，最終的に無音区間を含め全ての区間を結合する事により，処理制度の向上を図る．

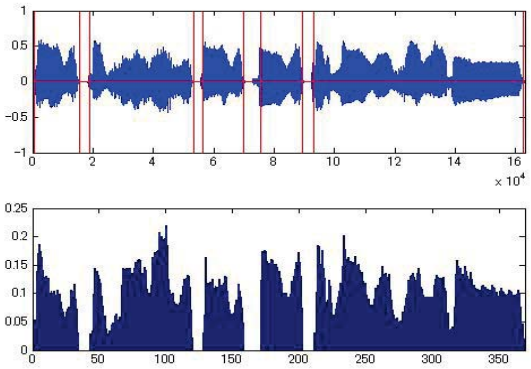

図 11　歌唱区間の割り出しの様子
Fig. 11　Extraction of singing section.

4.　聴取実験

3 章までに述べた二種の残差変換法，及び応用法によってハスキーボイス化した音声に対する聴取実験を行った．

実験の方法は，二つのハスキー化合成歌唱音声を聞き比

ⓒ 2013 Information Processing Society of Japan

べ，どちらが好みかを判断させる二点嗜好法を用いた．被験者は男性８名，女性４名の計１２名である．

聴取実験に使用した音声は

・ノイズ強調法
・間引き法
・イコライザー応用法
・自動生成法
・イコライザー応用法＋自動生成法

の五つである[f]．

聴取実験の集計結果は以下の図 12 のようになった．

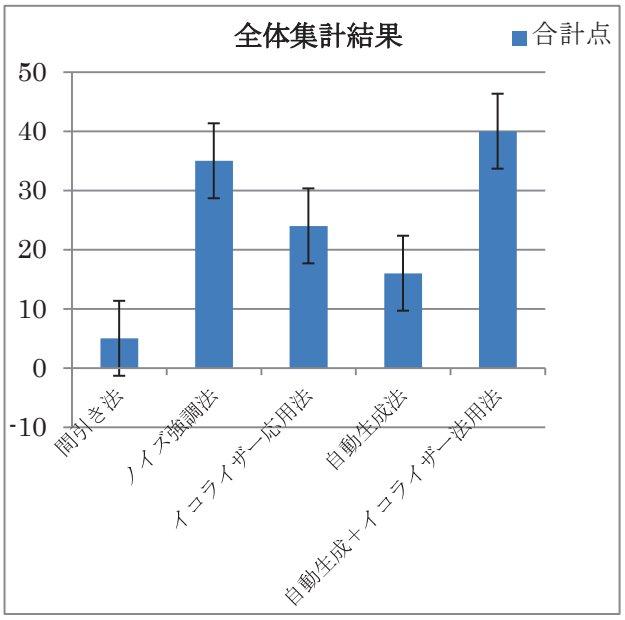

図 12　聴取実験の集計結果

Fig. 12　Aggregate results of the listening tests.

最も票を集めたのが，イコライザー応用法と自動生成法を掛け合わせたものであった．続いて，ノイズ強調法も大きな票を集めた．さらに各手法を対で比べた結果に注目していく．基本的な残差変換法については，ノイズ強調法に有用性があると確認された（図 13）．これは，ノイズ強調法が原音に含まれる音声を利用しているため，疑似ノイズを発生させる間引き法よりも，変換後の音声の自然性が保持されたためと推測される．そのため，全体集計でもノイズ強調法は大きな票を集めた．イコライザー応用法の有無の結果（図 14）では，有りに票が集まり，イコライザー応用法について有用性が確認された．自動生成法の有無の結果（図 15）は，多少検定を緩め，P≤0.15 とした場合に限り，有用性が確認された．結果的に二つの応用法，共に有用性があり，また全体集計で，イコライザー応用法＋自動生成法が最も票を集めた事から，両者の応用法に関して，重ね掛け相乗効果も期待できると判明した．全体集計の結果では，ノイズ強調法も多く票を集めたが，この手法はパラメーター設定を手動入力で行う必要性がある為，汎用性の面を考慮しても，イコライザー応用法＋自動生成法が最も有効なハスキーボイス化の手段であると考えられる．

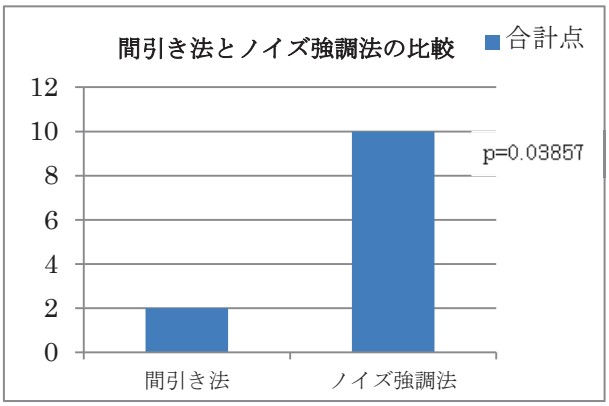

図 13　間引き法とノイズ強調法の比較

Fig. 13 Comparison of noise enhancement and thinning method.

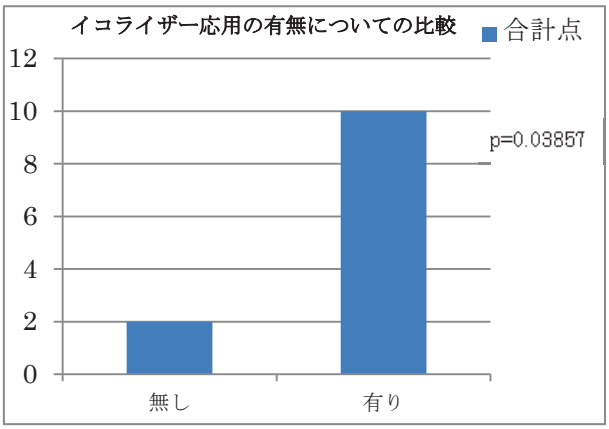

図 14　イコライザー応用の有無についての比較

Fig. 14　Comparison of the presence or absence of equalizer application.

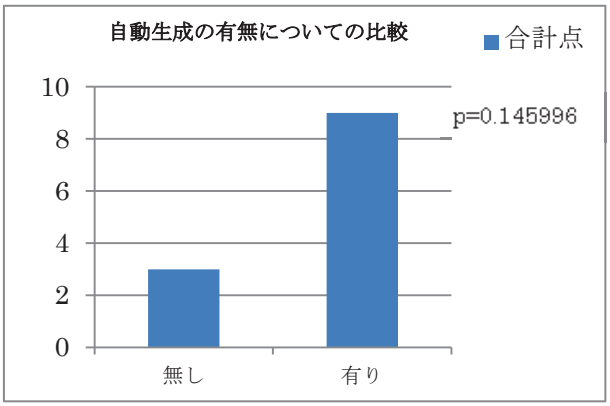

図 15　自動生成の有無についての比較

Fig. 15　Comparison for the presence of auto-generated.

[f]元音となる歌唱合成音声は，フリーソフト「ＵＴＡＵ」を使用して制作

ⓒ 2013 Information Processing Society of Japan

ここで，最も表を取ったイコライザー応用法＋自動生成法の周波数解析の結果を示しておく（図 16）．結果からわかるように周波数帯域を散布させる事に成功している．

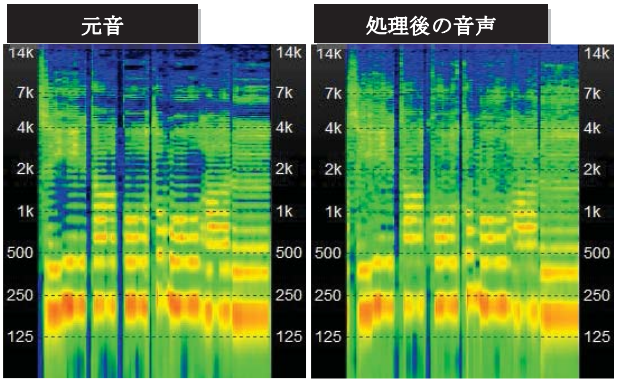

図 16　処理後の周波数解析結果

Figure 16　Frequency analysis results after treatment

5. おわりに

本研究では，歌唱合成音声のハスキーボイス化の方法について二つの残差変換法，及び応用法を提案し，それらの手法で生成した音声に対する聴取実験を行った．注目したいのは，この聴取実験で自動生成法に有用性が確認された事である．この自動生成法は，振幅のヒストグラムから算出される誤差αを元に処理レヴェルを決定している．すなわち，この誤差αはハスキーボイスの魅力を図る一つの指標であると考えられ，これは目標としていた人間的魅力を図る指標ともなり得る可能性がある．今後，さらに研究を深め，この人間的魅力を図る指標の確立を図りたい．

また，本研究では男性の合成音声のみを対象にハスキーボイス化を行ってきた為，女性の合成音声に関してもハスキーボイス化を可能とするのが第一の課題である．女性の合成音声のハスキーボイス化に際しては，男性のものと比べ周波数が高い点や，フォルマントの形状が異なる点を考慮した処理が必要であると推測される．

最後に，合成音声のハスキーボイス化の利用方法について記述する．これはもちろん，音楽制作への起用が第一として挙げられるが，それ以外にも，例えば，警備システムの警告音声など，相手に心理的影響を与えたい場合にも，このような特殊な合成音声が効果を発揮する可能性がある．すでに交通機関などのアナウンスで起用されているように，今後，社会の様々な場面での合成音声の使用が増加する事が予想される．そのため，合成音声に人間的魅力や特殊な効果を付加する技術の必要性は必ず，高まっていくであろう．そうなった時に，本研究の成果が貢献できるよう，これからも研究を続け，さらなる発見，進展に至れれば良いと考えている．

謝辞　研究にあたり，協力をしていただいた全ての皆様に感謝の意を表する．

参考文献

[1] 飴屋／菖蒲（あめや・あやめ）：歌声合成ツールＵＴＡＵサポートページ (http://utau2008.web.fc2.com/), Kenchan : UTAU のホームページ http://kenchan22.web.fc2.com/homepageright.html
[2] 徳田恵一，大浦圭一郎: 自動学習により人間のように歌う音声合成システム -Sinsy-，情報処理学会研究報告, (2012.2.3)
[3] 高久史麿，猿田享男，北村惣一郎：福井次矢: 家庭医学大全化，法研, (2010.10.10)
[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), No.14
[5] 加藤圭造，伊藤彰則: 複数歌唱者によるスクリーム歌唱音声の音響的特徴の分析，日本音響学会秋季講演論文集,1-2-9,(2012)
[6] 土師知行，楯敬藏，片岡英幸，藤木暢也: 残差波形分析による嗄声の客観的評価，耳鼻臨床 86: 7; 1013～1018, (1993.3.13)
[7] 青木直史：ディシタル・サウンド処理入門, CQ 出版社 (2006.04)
[8]　Ｌマゼル: 確率・統計・ランダム過程 (佐藤平八訳), 86～90頁，森北出版，東京, (1980)

ⓒ 2013 Information Processing Society of Japan