Vol.2019-MUS-123 No.54
Vol.2019-SLP-127 No.54
2019/6/23

# メルケプストラムを加工した音声の音質を評価する知覚モデルの開発

小川 樹[1,a)]　森勢 将雅[2]

**概要**：音声の加工は，計算機の性能の向上や小型端末の普及，音声研究の躍進により，誰でも手軽に行えるようになった．音声の加工には，音の 3 要素と呼ばれる「大きさ」,「高さ」,「音色」をそれぞれ加工する方法が広く用いられている．しかし，音色の加工は，大きさや高さの加工に対して加工に伴う劣化の予測が困難という問題点がある．そこで，音声の音色に特化した加工に伴う劣化を計測する知覚モデルを開発し，問題の解決を試みた．様々なスペクトル尺度と距離関数の組み合わせと音質の関係を調査し，その結果を用いて知覚モデルを開発した．

## 1. はじめに

音声の加工は，計算機の性能の向上や小型端末の普及，音声研究の躍進により，誰でも手軽に行えるようになった．音声を加工する機械やソフトウェアは，ボイスチェンジャーと呼ばれ，ロボットを想像させるような声や特定の芸能人に似た声など，さまざまな声に加工を行うことができる．この技術の基盤として，音の 3 要素と呼ばれる「大きさ」,「高さ」,「音色」をそれぞれ加工する方法が広く用いられている．しかし，音色の加工は，大きさや高さの加工に対して直感的な加工が困難という問題点がある．

大きさや高さは，加工するパラメータが 1 時刻あたり 1 次元の時系列であるが，音色は，加工するパラメータが 1 時刻あたり多次元なスペクトル包絡というスペクトル情報で表現されるためである．また，僅かな加工で音質が悪化することもあり，情報量の多さからその原因を特定することは難しい．そのため，目的の加工音声を作成するためには，人間が直接音声を聴取して評価を行い，再度加工を行うといった作業を繰り返し行う必要がある．この評価には，主観評価を行うことが最も正確であるが，大量の加工音声に対して行う場合，多くの時間を費やす必要があり，効率が悪い．そこで，主観評価の結果を推定する客観評価法を用いることが多い．PESQ (perceptual evaluation of speech quality) [1] と POLQA (perceptual objective listening quality assessment) [2] は，広く用いられている客観評価法である．しかし，PESQ は，電話帯域の音声を対象としており，POLQA は，音声の長さ等に制約があるため，任意の音声での評価は難しい．このことから，任意の音声に対しても，音色の加工を行った音声を評価することができる知覚モデルの構築を目指す．

本研究では，音声の音色の加工に伴う劣化の予測が困難という問題点を解決するため，音色の加工後に起こる音質劣化に特化した知覚モデルを開発し．既存の音声の客観評価法と音色を表す音響特徴量の関係を調査する．調査した結果より，複数の知覚モデルを開発する．

## 2. 音声の評価法

音声を加工することにより，定常的に起きるノイズや，局所的に生じる振幅のピークなどの劣化が生じることがある．この音声の音質劣化を評価する方法がいくつか存在する．MOS (mean opinion score) 評価は主観評価法であり，音質について「非常に良い」から「非常に悪い」までの 5 段階で評価する方法である．しかし，評価者や評価音声の準備にコストがかかるという問題点がある．評価者には，正常な聴取能力を持つ複数人を集める必要があり，音声によっては事前の練習が必要になる．また，評価環境は，専用の評価施設で行い，騒音や音圧レベルなどの条件を揃えたり，評価音声の順序を毎回変化させたりするなど，評価に対するばらつきを抑える必要がある．この問題点を解決し，MOS 評価と同等の評価値を推定する方法として，PESQ と POLQA や，AutoMOS [3] などの客観評価法が提案されている．

PESQ と POLQA は，参照音声と評価音声を比較し，知覚・認知モデルのそれぞれの処理から，評価値を推定する方法である．PESQ は電話帯域の音声を対象としており，

[1] 山梨大学
[2] 明治大学
[a)] g19tk006@yamanashi.ac.jp

POLQA は PESQ を拡張し，性能の向上に加え，より広帯域な音声も対象としている．PESQ と POLQA は，共に国際規格となっている．しかし，PESQ はサンプリング周波数が 16 kHz を上回る音声に対応しておらず，POLQA は使用する音声への制約，特に時間に関する制約が多い．サンプリング周波数 48 kHz に対応させた PESQ の拡張版である EW-PESQ [4] も提案されているが，音色の加工を行った音声の評価法としての検討が十分であるとは言い難い．近年では，AutoMOS と呼ばれる，ニューラルネットワークを用いた客観評価法も存在する．AutoMOS は，スペクトルとその動的特徴量を入力として，評価値を出力するように学習を行う．text-to-speech システムなどで合成された音声や，評価音声しか用意できない環境に対しても利用することができ，その利用範囲は広い．しかし，学習を行うためのデータセットが，文献中では約 17 万音声と非常に多く，同規模のデータセットを用意することは困難である．

上記の評価法は，音声を対象にしていたが，スペクトル包絡のみを評価する指標として，スペクトル距離がある．スペクトル距離は，参照音声と評価音声のそれぞれのスペクトル包絡を，適当な距離関数を用いて得られた誤差の尺度であり，2 つのスペクトル包絡の全周波数・全時間に対する積分値又はその平均で求めることができる．

## 3. 提案法

### 3.1 本研究の位置づけ

本研究では，第 2 節で述べた PESQ と EW-PESQ を従来法とする．POLQA は，今回使用する音声に適していないため，また，AutoMOS は，学習するためのデータセットに結果が依存し，再現性を担保することができないため，今回比較する従来法から除外した．音色のみを加工した音声の音質劣化の推定を目的とするため，音色のみの加工を行った音声を評価音声とし，加工を行う前の音声を参照音声として，2 つの音声を比較する評価法とする．サンプリング周波数が 40 kHz 以上のフルバンド音声も対象とし，信号処理のみを用いて評価値を推定する．評価値は，PESQ と同様に，主観評価法の評価値と同等の評価値を推定する．

### 3.2 メルケプストラムを用いた音声変換

音色のみの加工を行うため，音声から音色を表す音響特徴量を抽出する．音響特徴量の抽出には，音声分析合成システムである WORLD [5] を用いた．WORLD は，高品質な音声の分析合成システムであり，声の高さを表す基本周波数，声の音色を表すスペクトル包絡，声のかすれ具合を表す非周期性指標の 3 つのパラメータを音声から抽出する．これらの 3 つのパラメータを適宜変更し，WORLD で合成することで，音質劣化の少ない合成音声を生成する．同じ音声データセットを用いて，声質変換 [6] の精度を競う Voice Conversion Challenge 2016 [7] では，参加した 17 チームのうち，13 チームでメルケプストラム [8] を用いたシステムを開発しており，性能の面でもメルケプストラムを用いたシステムが上位に多いこと [9] から，メルケプストラムを用いた音声変換を行う．メルケプストラムとは，人間の聴覚特性に合わせて，高次元のスペクトル包絡を，低次元の特徴量に圧縮したものである．加工方法は，メルケプストラム次数の 1 から $N$ 次までの 1 刻みのうち，一つの次元のみを $-8$ から 10 まで 1 刻みで定数倍する．加工したメルケプストラムをスペクトル包絡に復元し，保存しておいた基本周波数と非周期性指標と共に WORLD に入力，合成することで音声変換を行う．

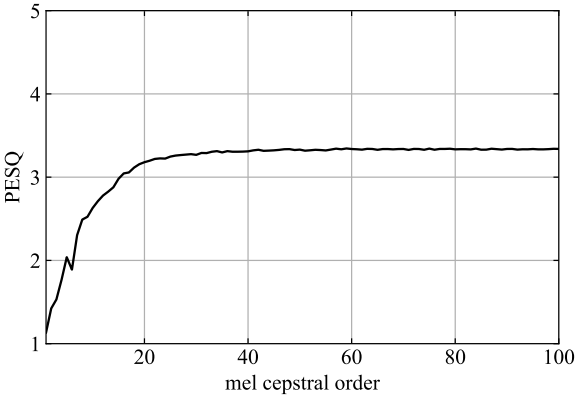


**図 1** メルケプストラム次数の変化による PESQ の評価値．

メルケプストラム次数 $N$ を決めるために，メルケプストラム次数を 1 から 100 次にした際の，PESQ の評価値の変化について予備調査を行った．使用音声は，JSUT (Japanese speech corpus of Saruwatari laboratory, the University of Tokyo) [10] から女性話者の 10 音声，HTS-demo_NIT-ATR503-M001 [11] から男性話者の 10 音声，計 20 音声を用いた．調査結果を，図 1 に示す．縦軸は PESQ の評価値，横軸はメルケプストラム次数を表す．図より，28 次以上で PESQ の評価値の変動が鈍くなったため，28 次を音声変換に用いるメルケプストラム次数と決定した．

### 3.3 スペクトルの種類

調査するスペクトルとして，WORLD のスペクトル包絡に加え，以下の項で述べる 5 種類のスペクトルを用いる．

#### 3.3.1 メル尺度

メル尺度 (mel) [12] は，音の高さの知覚的尺度である．低周波数の高さには敏感だが，高周波数の高さには鈍感という聴覚特性を元に作成された．周波数からメル尺度への変換は，文献 [13] から，式 (1) を用いる．

$$\mathrm{mel}(f) = 1127.01048 \log\left(\frac{f}{700} + 1\right) \qquad (1)$$

$f$ は周波数を表す．式 (1) 以外にもいくつかの変換式が提案されているが，どの変換式も低周波数では線形，高周波

数では対数の形を模している．本研究では，WORLD で得られたスペクトル包絡に，100 次のメルフィルタバンクをかけたものをメルスペクトルとして用いる．フィルタバンクとは，特定の周波数軸上で等間隔なバンドパスフィルタである．

#### 3.3.2 バーク尺度

バーク尺度 (Bark) [14] は，Zwicker によって提案された音響心理学的尺度である．臨界帯域幅測定法を用いた心理学実験を元に作成された．周波数からバーク尺度への変換は，文献 [15] から，式 (2) を用いる．

$$\mathrm{Bark}(f) = \frac{26.81f}{1960+f} - 0.53 \tag{2}$$

$f$ は周波数を表す．バーク尺度も，メル尺度と同様に，いくつかの変換式が提案されている．本研究では，WORLD で得られたスペクトル包絡に，100 次のバークフィルタバンクをかけたものをバークスペクトルとして用いる．

#### 3.3.3 ERB 尺度

ERB (equivalent rectangular bandwidth) 尺度 [16] は，Moore らによって提案された音響心理学的尺度である．バーク尺度で利用されていた臨界帯域幅測定法を改良した，ノッチ雑音マスキング法を用いた心理学実験を元に作成された．周波数から ERB 尺度への変換は，文献 [17] から，式 (3) を用いる．

$$\mathrm{ERB}(f) = 21.4\log_{10}\left(\frac{4.37f}{1000}+1\right) \tag{3}$$

$f$ は周波数を表す．本研究では，音声波形から得られた 100 次の ERB スペクトルを用いる．

#### 3.3.4 ガンマチャープ

ガンマチャープ [18] は，音圧のレベル依存性や圧縮特性といった聴覚末梢系の非線形性や，時間変化による動的な特性をモデル化した聴覚フィルタである．このガンマチャープには，線形で時不変なガンマチャープ (gammachirp: GC)，非線形で時不変な圧縮型ガンマチャープ (compressive gammachirp: cGC)，非線形で時変な動的圧縮型ガンマチャープの 3 種類がある．聴覚特性を最もよく表現しているフィルタは，動的圧縮型ガンマチャープであるが，非線形性があるため波形に対する音圧レベルという他の尺度には存在しないパラメータが必要であること，及び，時変性を取り入れるため計算に時間がかかることから，本研究では，GC と cGC のみを用いる．

### 3.4 距離関数の種類

調査する距離関数として，以下で述べる 5 種類の距離関数と，それらに対数をとった距離関数を加えた計 10 種類を用いる．

$$D_{\mathrm{EU}} = \frac{1}{T}\int_0^T \sqrt{\frac{1}{f_{\mathrm{N}}}\int_0^{f_{\mathrm{N}}} \left(P(t,f)-\hat{P}(t,f)\right)^2 df}\,dt \tag{4}$$

式 (4) は，ユークリッド距離である．$P(t,f)$ は，真値のスペクトル包絡の時間周波数表現であり，$\hat{P}(t,f)$ は，加工したスペクトル包絡の時間周波数表現である．$T$ は信号長に相当し，$t$ は分析時刻を，$f_{\mathrm{N}}$ はナイキスト周波数であり，$f$ は周波数を示す．この距離関数は，単純な尺度であり，誤差の正負に関わらず対称である．

$$D_{\mathrm{LS}} = \frac{1}{T}\int_0^T \sqrt{\frac{1}{f_{\mathrm{N}}}\int_0^{f_{\mathrm{N}}} \left(10\log_{10}\left(\frac{P(t,f)}{\hat{P}(t,f)}\right)\right)^2 df}\,dt \tag{5}$$

式 (5) は，対数スペクトル距離である．ユークリッド距離を，対数軸上で比較した距離関数となる．

$$D_{\mathrm{IS}} = \frac{1}{T}\int_0^T \frac{1}{f_{\mathrm{N}}}\int_0^{f_{\mathrm{N}}} D_{\mathrm{IS}}(t,f)\,df\,dt \tag{6}$$
$$D_{\mathrm{IS}}(t,f) = \frac{P(t,f)}{\hat{P}(t,f)} - \log\left(\frac{P(t,f)}{\hat{P}(t,f)}\right) - 1$$

式 (6) は，板倉斉藤距離 [19] である．ユークリッド距離では対称であった正負に対して，負方向には大きく，正方向には小さく距離を取る，非対称性を持つ．スペクトル包絡のピークが弱まるより強まる加工が音声として自然なため，この距離関数は音声に適していると言える．

$$D_{\mathrm{WIS}} = \frac{1}{T}\int_0^T \frac{1}{0.45f_s - 2f_0}\int_{2f_0}^{0.45f_s} D_{\mathrm{WIS}}(t,f)\,df\,dt \tag{7}$$
$$D_{\mathrm{WIS}}(t,f) = \left(\frac{P(t,f)}{\hat{P}(t,f)} - \log\left(\frac{P(t,f)}{\hat{P}(t,f)}\right) - 1\right)u(f)$$
$$u(f) = \frac{9.294}{0.00437f+1} \tag{8}$$

式 (7) は，重み付き板倉斎藤距離である．$f_s$ はサンプリング周波数を，$f_0$ は基本周波数を示す．板倉斎藤距離に，低域ほど大きく，高域ほど小さい周波数重みをかけ合わせたもので，周波数重みは，式 (3) の導関数として，式 (8) で表される．積分範囲は，低域は重みが大きくなりすぎるため，高域は折り返しの影響を除くために狭くしている．

$$D_{\mathrm{dB}} = \sqrt{\frac{1}{T}\int_0^T \frac{1}{f_{\mathrm{N}}}\int_0^{f_{\mathrm{N}}} D_{\mathrm{dB}}(t,f)\,df\,dt} \tag{9}$$
$$D_{\mathrm{dB}}(t,f) = \left(10\log_{10}\left(\frac{P(t,f)}{\bar{P}(t)}\right) - 10\log_{10}\left(\frac{\hat{P}(t,f)}{\bar{\hat{P}}(t)}\right)\right)^2$$

式 (9) は，文献 [20] で提案された距離関数である．$\bar{P}(t)$ は，真値のスペクトル包絡の周波数の平均値であり，$\bar{\hat{P}}(t)$ は，加工したスペクトル包絡の周波数の平均値である．文献中では，ガンマチャープを用いた声道長の推定のために用いられている．

## 4. 従来法と提案法の比較

### 4.1 実験条件

知覚モデルを開発するための予備実験として，従来法と

Vol.2019-MUS-123 No.54
Vol.2019-SLP-127 No.54
2019/6/23

表 1　従来法と提案法の比較実験の実験条件.

| | |
|---|---|
| メルケプストラム次数 | 28 次 |
| 変化量 | 1 刻みで −8 から 10 倍の 19 通り |
| 使用音声 | FW07 |
| 音源 | 全 40 音声 (男女各 2 名 × 10 文章) |
| 加工種類 | 全 532 種類 (28 次 × 19 通り) |
| 従来法 | PESQ, EW-PESQ |
| 提案法 | 全 60 種類<br>(スペクトル 6 種類 × 距離関数 10 種類) |

表 2　従来法と提案法の距離との相関係数.

| 従来法 | 提案法 | 相関係数 |
|---|---|---|
| PESQ | ERB-log $(D_{\mathrm{dB}})$ | −0.814 |
| | GC-log $(D_{\mathrm{LS}})$ | −0.801 |
| | Bark-log $(D_{\mathrm{dB}})$ | −0.792 |
| EW-PESQ | ERB-log $(D_{\mathrm{dB}})$ | −0.753 |
| | GC-log $(D_{\mathrm{LS}})$ | −0.746 |
| | GC-log $(D_{\mathrm{dB}})$ | −0.724 |

提案法の比較実験を行い，従来法と提案法の相関について調査する．実験条件を，表 1 に示す．使用音声は，親密度別単語了解度試験用音声データセット 2007 (familiarity-controlled wordlists 2007: FW07) [21] を用いた．この音声データセットは，サンプリング周波数が 48 kHz であり，PESQ は 16 kHz までの音声にしか対応しておらず，そのままのサンプリング周波数では評価が行えない．そのため，PESQ で評価を行う際は，音声を 16 kHz にダウンサンプリングする．

### 4.2　実験結果

まず，PESQ 又は EW-PESQ と提案法の距離との相関係数のうち，それぞれ上位 3 つについて表 2 に示す．次に，従来法のそれぞれで最も相関係数が高かった組み合わせについて，散布図と回帰直線を図 2, 3 に示す．縦軸は PESQ 又は EW-PESQ の評価値，横軸は提案法の距離を表し，右上に相関係数を示す．

距離関数では, 対数軸上で比較する $\log(D_{\mathrm{LS}})$ や $\log(D_{\mathrm{dB}})$ を用いた提案法が，従来法との相関が強かった．対数軸上で比較することにより，線形軸上で比較するより距離の範囲が狭まるためだと考えられる．また，従来法と提案法の間に非線形性があることも確認できる．

## 5.　知覚モデルの開発

### 5.1　非線形モデルの推定

スペクトル距離から評価値を推定する知覚モデルを開発する．知覚モデルには，4.2 項より従来法の評価値と提案法の距離との間に非線形性が見られたため，非線形モデル

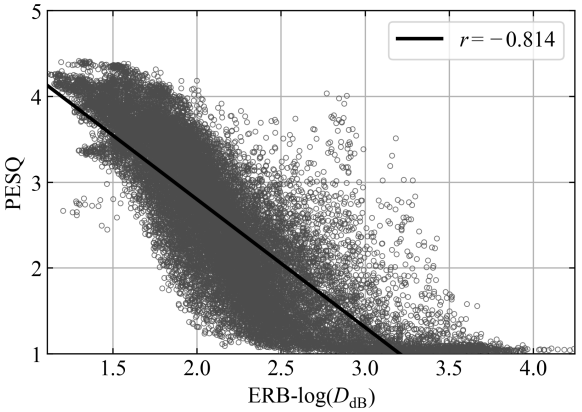


図 2　PESQ と ERB-log $(D_{\mathrm{dB}})$ の距離との散布図と回帰直線.

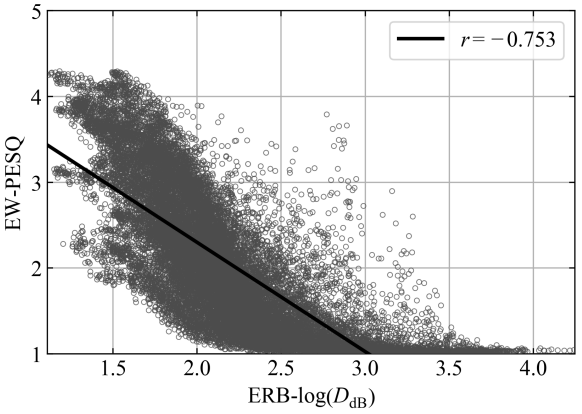


図 3　EW-PESQ と ERB-log $(D_{\mathrm{dB}})$ の距離との散布図と回帰直線.

を使用する．非線形のモデルとして， [22] から，以下の 3 種類の関数を用いる．

$$\text{Exp} \qquad y = a\mathrm{e}^{bx} + c \tag{10}$$

$$\text{Shah} \qquad y = a + bx + cd^{x} \tag{11}$$

$$\text{Stirling} \qquad y = a + b\frac{\mathrm{e}^{cx} - 1}{c} \tag{12}$$

$x$ はスペクトル距離，$y$ は評価値，$a, b, c, d$ は各非線形モデルのパラメータである．パラメータは，4.2 項のデータを基に，レーベンバーグ・マーカート法 [23] によって求める．

### 5.2　開発した知覚モデルと従来法の比較

開発した 360 種類の知覚モデルと従来法の比較を行う．PESQ 又は EW-PESQ と開発した知覚モデルとの相関係数のうち，それぞれ上位 3 つについて表 3 に示す．次に，従来法のそれぞれで最も相関係数が高かった組み合わせについて，散布図と回帰直線を図 4, 5 に示す．縦軸は PESQ 又は EW-PESQ の評価値，横軸は提案法の評価値を表し，左上に相関係数を示す．

PESQ では，WORLD-log $(D_{\mathrm{IS}})$ や ERB-$D_{\mathrm{dB}}$ の組み合わせが，EW-PESQ では，ERB-$D_{\mathrm{dB}}$ や ERB-log $(D_{\mathrm{dB}})$ の組み合わせが，従来法と相関の強い知覚モデルとなった．また，非線形モデルの種類については，大きな差は見られ

表 3　従来法と知覚モデルの評価値との相関係数.

| 従来法 | 提案法 | 相関係数 |
|---|---|---|
| PESQ | WORLD-log ($D_{\mathrm{IS}}$)-Shah | 0.851 |
| | ERB-$D_{\mathrm{dB}}$-Shah | 0.848 |
| | ERB-$D_{\mathrm{dB}}$-Exp | 0.848 |
| EW-PESQ | ERB-$D_{\mathrm{dB}}$-Shah | 0.814 |
| | ERB-$D_{\mathrm{dB}}$-Exp | 0.814 |
| | ERB-$D_{\mathrm{dB}}$-Stirling | 0.814 |

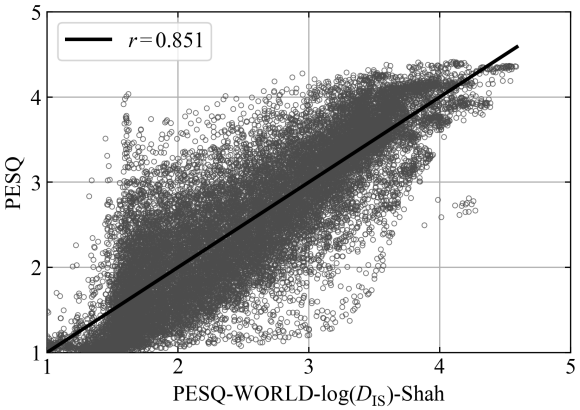


図 4　PESQ と PESQ-WORLD-log ($D_{\mathrm{IS}}$)-Shah の評価値との散布図と回帰直線.

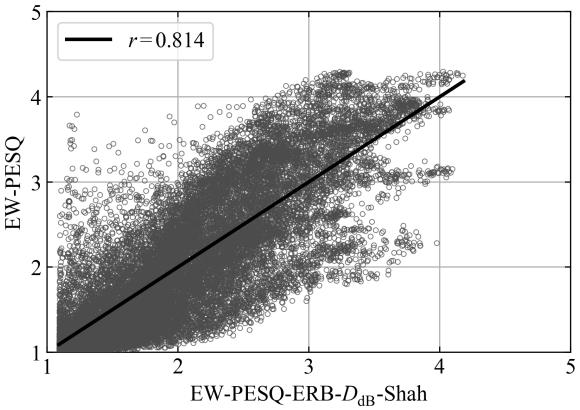


図 5　EW-PESQ と EW-PESQ-ERB-$D_{\mathrm{dB}}$-Shah の評価値との散布図と回帰直線.

なかったが，Stirling のモデルは，いくつかの知覚モデルに対して，パラメータは収束しているが曲線の当てはめに失敗していることがあった．最も従来法との相関が強かった PESQ-WORLD-log ($D_{\mathrm{IS}}$)-Shah のモデルを，式 (13) に示す．

$$y = 1.68376 - 0.01777x + 1.50614 \times 0.50027^{x} \quad (13)$$

## 6. 考察

まず，本研究では，相関係数に基づいて開発した．2 つの提案法を，相関係数の値のままで比較することはできないが，2 つの相関係数の差を比較する方法として，文献 [24] がある．今回は，提案法同士の比較は，必要ではないと考えて行っていないが，今後行う予定の主観評価実験の結果を用いた従来法と提案法の相関係数の比較を行う際には，使用したいと考えている．

次に，本研究では，非線形モデルとして 3 種類のみを用いた．この 3 種類以外の非線形モデルや線形モデルも利用して知覚モデルの推定を行ったが，その中で首尾よく当てはめが行えた 3 種類のみを選定した．また，線形モデルより非線形モデルが適しているのは，図 2, 3 から分かる通り，従来法と提案法の距離の間に，非線形な関係性があることが確認できた．そのため，複数の非線形モデルを対象とした知覚モデルを作成し，より相関が高くなるように構築を行った．

最後に，開発した知覚モデルは，従来法のデータを用いている．そのため，新たに実施した主観評価実験のデータを反映させているわけではない．しかし，従来法である PESQ や EW-PESQ は，ばらつきが大きく，このばらつきを抑える目的として，非線形モデルの当てはめによる新たな知覚モデルの開発は，妥当であると考えられる．

## 7. おわりに

本研究では，音声の音色の加工に伴う劣化の予測を行うための，知覚モデルを開発した．提案法として，音声の音色を表すスペクトルと，スペクトルを評価する距離関数に着目し，スペクトルを 6 種類，距離関数を 10 種類の全 60 種類について検討した．まず，従来法と提案法の距離との比較を行い，従来法である PESQ と EW-PESQ では，PESQ がより提案法との相関が強い結果となった．さらに，距離関数については，対数軸上で比較する距離関数の相関が強くなることも確認した．次に，従来法と提案法の比較実験の結果を基に，知覚モデルの開発を行った．実験結果より，従来法と提案法の距離との間には非線形な関係性があることから，非線形モデルをベースとした知覚モデルの推定を行った．従来法と開発した知覚モデルの性能の評価を行い，WORLD のスペクトルや ERB 尺度と対数軸上で比較する距離関数を組み合わせた知覚モデルが，従来法との相関が強い結果となった．

今後の課題として，まず，主観評価実験を行う．本研究で開発した知覚モデルが，従来法よりも優れているかについて評価するためである．また，主観評価実験の結果に最も近い知覚モデルを，最適な知覚モデルとして決めるためにも必要である．他にも，より複雑な音色の加工を行った音声への性能評価も挙げられる．今回の評価音声の加工方法は，1 つに限定しているため，他の方法を用いて音色を加工した音声に対しての評価値が，正しく推定できるとは限らない．そのため，より多様な加工を行った音声に対しても頑強であるかの評価は必要である．

**謝辞**　ガンマチャープについて御教授を賜った和歌山大学システム工学部システム工学科入野俊夫教授に深謝す

る．本研究は，科研費 JP16H05899, JP16H01734，JST さきがけ JPMJPR18J8 の支援を受けた．

## 参考文献

[1] ITU-T: Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, Recommendation P.862, International Telecommunication Union (2001).

[2] ITU-T: Perceptual objective listening quality prediction, Recommendation P.863, International Telecommunication Union (2011).

[3] Patton, B., Agiomyrgiannakis, Y., Terry, M., Wilson, K., Saurous, R. A. and Sculley, D.: AutoMOS: learning a non-intrusive assessor of naturalness-of-speech, *NIPS 2016 End-to-end Learning for Speech and Audio Processing Workshop* (2016).

[4] Bispo, B. C., Esquef, P. A. A., Biscainho, L. W. P., Lima, A. A. d., Freeland, F. P., Jesus, R. A. d., Said, A., Lee, B., Schafer, R. W. and Kalker, T.: EW-PESQ: a quality assessment method for speech signals sampled at 48 kHz, *Journal of the Audio Engineering Society*, Vol. 58, No. 4, pp. 251–268 (2010).

[5] Morise, M., Yokomori, F. and Ozawa, K.: WORLD: a vocoder-based high-quality speech synthesis system for real-time applications, *IEICE Transactions on Information and Systems*, Vol. E99.D, No. 7, pp. 1877–1884 (2016).

[6] Stylianou, Y., Cappe, O. and Moulines, E.: Continuous probabilistic transform for voice conversion, *IEEE Transactions on Speech and Audio Processing*, Vol. 6, No. 2, pp. 131–142 (1998).

[7] Toda, T., Chen, L.-H., Saito, D., Villavicencio, F., Wester, M., Wu, Z. and Yamagishi, J.: The voice conversion challenge 2016, *Interspeech 2016*, pp. 1632–1636 (2016).

[8] Fukada, T., Tokuda, K., Kobayashi, T. and Imai, S.: An adaptive algorithm for mel-cepstral analysis of speech, *ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 137–140 (1992).

[9] Wester, M., Wu, Z. and Yamagishi, J.: Multidimensional scaling of systems in the voice conversion challenge 2016, *9th ISCA Speech Synthesis Workshop*, pp. 38–43 (2016).

[10] Sonobe, R., Takamichi, S. and Saruwatari, H.: JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis, arXiv preprint, 1711.00354 (2017).

[11] HTS Working Group: The NITech Japanese speech database NIT ATR503 M001 (2019-02-15).

[12] Stevens, S. S., Volkmann, J. and Newman, E. B.: A scale for the measurement of the psychological magnitude pitch, *Journal of the Acoustical Society of America*, Vol. 8, No. 3, pp. 185–190 (1937).

[13] Stevens, S. S. and Volkmann, J.: The relation of pitch to frequency: a revised scale, *The American Journal of Psychology*, Vol. 53, No. 3, pp. 329–353 (1940).

[14] Zwicker, E.: Subdivision of the audible frequency range into critical bands (Frequenzgruppen), *Journal of the Acoustical Society of America*, Vol. 33, No. 2, p. 248 (1961).

[15] Traunmüller, H.: Analytical expressions for the tonotopic sensory scale, *Journal of the Acoustical Society of America*, Vol. 88, No. 1, pp. 97–100 (1990).

[16] Moore, B. C. J. and Glasberg, B. R.: Suggested formulae for calculating auditory-filter bandwidths and excitation patterns, *Journal of the Acoustical Society of America*, Vol. 74, No. 3, pp. 750–753 (1983).

[17] Glasberg, B. R. and Moore, B. C. J.: Derivation of auditory filter shapes from notched-noise data, *Hearing Research*, Vol. 47, No. 1, pp. 103–138 (1990).

[18] Irino, T. and Patterson, R. D.: A dynamic compressive gammachirp auditory filterbank, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No. 6, pp. 2222–2232 (2006).

[19] Chan, A. H.-S. and Ao, S.-I.: *Advances in industrial engineering and operations research*, Springer (2008).

[20] 入野俊夫，河原英紀，Patterson, R. D.：聴覚におけるスケール分析のための末梢系フィルタバンクのウェーブレット性と非線形性，数理解析研究所講究録，Vol. 1928, pp. 27–57 (2014).

[21] 近藤公久，天野成昭，坂本修一，鈴木陽一：親密度別単語了解度試験用音声データセット 2007(FW07)，NII 音声資源コンソーシアム (2007).

[22] Lightstone®: 非線形フィット関数の一覧 | データ分析・グラフ作成 Origin| ライトストーン (2019-01-31).

[23] Levenberg, K.: A method for the solution of certain non-linear problems in least squares, *Quarterly of Applied Mathematics*, Vol. 2, No. 2, pp. 164–168 (1944).

[24] 池田　央: 統計ガイドブック，新曜社 (1989).