Vol.2015-AVM-88 No.5
2015/2/27

視覚的テクスチャを表現するオノマトペの音響特徴量

三輪 篤志[1,a)]　坂東 敏博[1,b)]　佐々木 康成[2,c)]

概要：ヒトはオノマトペ（擬音語，擬態語の総称）を用いてテクスチャ（画像において色彩，濃淡の分布が形づくる表面の性質）を表現することがある．タイルの表面を「つるつる」，雲の質感を「もくもく」と表現する振る舞いは日常茶飯事的に行われている．本研究ではこれを視覚情報から音声情報への変換と捉え，その関係を調べた．さらにそこから，重要だとわかった音響特徴量，メル周波数ケプストラム係数の 8 次，12 次を用いて，音源に効果を付与するフィルタを設計した．

キーワード：質感，擬態語，画像解析，音声解析，主成分分析

Sound Feature Values of Onomatopoeia that Express Visual Texture

Miwa Atsushi[1,a)]　Bando Toshihiro[1,b)]　Sasaki Yasunari[2,c)]

Abstract: Human often express a texture by using onomatopoeia. It is natural behavior that they express a surface of tile as “tsurutsuru” or a texture of crowd as “mokumoku” in Japanese. We interpret this behavior as a conversion of visual information to auditory information and invested the relation between them. This study shows that 8th and 12th of mel-frequency Cepstrum coefficient are the most important feature values to express a texture as a result. Furthermore the filter that gives an effect of 8th and 12th of mel-frequency Cepstrum coefficient is designed in this report.

Keywords: Texture, Onomatopoeia, Image Analysis, Voice Analysis, PCA

1. はじめに

我々が使用している擬音語，擬態語などはオノマトペと呼ばれる．このオノマトペを利用するタイミングは日常的に多く見られるだろう．例えばきれいに磨かれた車のボディは「ピカピカ」だし，きめ細かく泡立った生クリームは「ふわふわ」に見えるかもしれない．

このときヒトはどのようにしてオノマトペを思いついて対象を表現しているのだろうか，という疑問の発想がまず本研究の発端である．“きれいに磨かれた車のボディ”や “きめ細かく泡立った生クリーム”といった物体の質感は総じてテクスチャ [*1] と呼ばれる．本研究ではヒトがテクスチャをみたとき，その視覚情報をもとに適切なオノマトペを選出しているのだと考えた．

オノマトペという音声情報を用いてテクスチャという視覚情報を表現するヒトの振る舞いは大変興味深い．本研究はこの振る舞いを計算機上で実現しようという動機のもと行った．このシステムが構築されれば，テクスチャ画像を音で聴き，今まで気づき得なかった異なる側面からの価値を発見できる可能性もあるだろう．

音を画像に変換しようという試みは既存の音楽プレーヤーなどでみられるが，その逆，画像から音への変換は未だ挑戦された例も少なく，多くの可能性を秘めている．

[1] 同志社大学大学院理工学研究科情報工学専攻
Graduate School of Science and Engineering, Doshisha University, 1-3 Tatara Miyakodani, Kyotanabe-shi, Kyoto-fu 610-0394, Japan
[2] 金沢星稜大学教養教育部
Kanazawa Seiryo University, 10-1 Ushi, Gosho-machi, Kanazawa-shi, Ishikawa Prefecture 920-8620, Japan
[a)] atsushimiwa.info@gmail.com
[b)] tbando@mail.doshisha.ac.jp
[c)] ysasaki@seiryo-u.ac.jp

[*1] 厳密には，画像として色彩，濃淡の分布が形作る表面の性質と定義できる．

画像から音への変換という応用的な目標には，解決すべき基礎的な問題が様々あるだろう．本研究ではその 1 つとして，テクスチャの画像としての情報とオノマトペの音としての情報の関係について詳しく調べた．

具体的には，まずどのようなテクスチャをみたとき，ヒトは種々のオノマトペを思いつくのか調べた．次にこの実験に用いたテクスチャ画像を解析し，様々な特徴量を算出した後，主成分分析した．この主成分とオノマトペの音としての特徴量の関連性を距離分析によって求め，どのようなオノマトペの情報がテクスチャを表現するのに重要なのかを調べた．

2. テクスチャから思いつくオノマトペ

ヒトがどのようなテクスチャをみたとき，どのようなオノマトペを思いつくのかを調べるために実験を行った．

2.1 実験方法

被験者 60 人（男性 30 人，女性 30 人）に対して 50 のテクスチャ画像をランダム順に見せ，それぞれの画像に対して 1 つ，思いついたオノマトペを回答させた．まずそれぞれ自由にオノマトペを回答させた（実験 A）．被験者によってはあるテクスチャのオノマトペを思いつかない場合があるため，50 のうち 10 以上のテクスチャへの回答を義務づけてこれに対応した．次に同じテクスチャ群に対して，用意した 102 のオノマトペから選択して回答させた（実験 B）．このときは 50 のテクスチャすべてへの回答を義務づけた．102 のオノマトペは「暮らしのことば擬音・擬態語辞典」[1] からの引用した．これらはすべて 4 モーラ [*2] に統一した．

被験者に提示するテクスチャは 1024 × 1024pixel に統一し，ディスプレイとの距離は 50cm とした．得られた回答を音素 [*3] に分解し，テクスチャごとに集計した後，その発現確率を音素の発現順に求めた．つまりあるテクスチャで選ばれたオノマトペにおける音素の発現順 i において，任意の音素の種が選ばれた頻度を f_{ph} としたとき，式 1 で表される $\boldsymbol{P_i}$ である．

$$\boldsymbol{P_i} = \frac{f_{ph}}{\sum_{ph} f_{ph}} \tag{1}$$

2.2 実験結果

図 1 のテクスチャ例の分析結果を示したのが図 2〜5 である．それぞれの音素が発現した確率を示した．第 1 子音では/s/が約 0.27，/t/が約 0.18，第 2 子音では/r/が約 0.46 の割合で発現した．第 1 母音では/u/が約 0.41，第 2 母音では/a/が約 0.68 の割合で発現した．本実験で扱ったオノ

[*2] ことばを数える単位の 1 つで「拍」とも言う．例えば「りんご」，「アップル」，「コンピュータ」はそれぞれ 3，4，5 モーラである．
[*3] 広義に音韻学上の最小単位を指す．本研究では/shy/，/chy/，/ny/，/zy/も区別して考えた．

図 1 実験で利用したテクスチャ画像例
Fig. 1 A texture example used in experiment

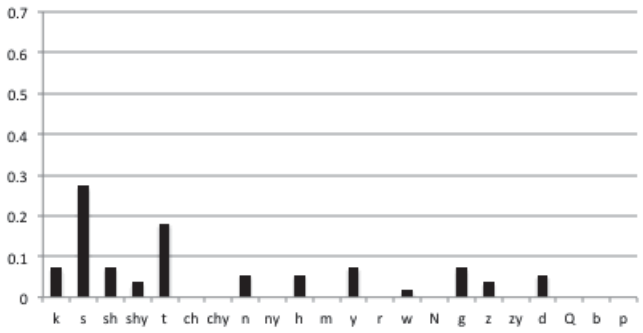

図 2 第 1 子音の発現確率
Fig. 2 Appearance probability of each consonant at 1st phoneme

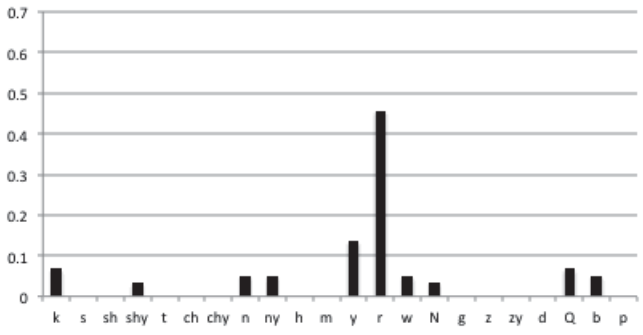

図 3 第 2 子音の発現確率
Fig. 3 Appearance probability of each consonant at 3rd phoneme

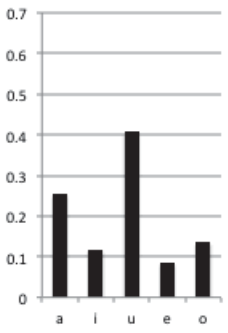

図 4 第 1 母音の発現確率
Fig. 4 Appearance probability of each vowel at 2nd phoneme

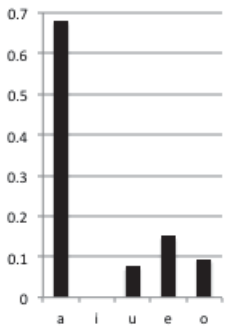

図 5 第 2 母音の発現確率
Fig. 5 Appearance probability of each vowel at 4th phoneme

マトペはほとんどが CVCV_CVCV[*4] の構成であったため，ここではその前半の結果のみ図示した．なお母音から始まるオノマトペに関しては第 1 子音は数えず，第 1 母音，第 2 子音，第 2 母音として数えた．

3. テクスチャ画像解析

実験で用いた 50 のテクスチャ画像の特徴量を解析した．

3.1 RGB，HSV 統計量

画像は計算機上で RGB それぞれ 256 階調のデータとして格納されて扱われている．任意の画像における RGB それぞれの平均と標準偏差を特徴量とした．また RGB を HSV に変換した後，これらの平均と標準偏差も特徴量とした．

3.2 濃淡ヒストグラムから得た特徴量

画像をグレースケール化し，濃淡ヒストグラムを求め，その平均，分散，歪度，尖度を求めた．なお濃淡ヒストグラムの全体が 1.0 になるよう正規化してから算出した．歪度は濃淡の分布形状がどれだけ対照形から歪んでいるか，尖度は分布が平均値のどれだけ近傍にあるかを示す特徴量である [2]．

3.3 修正局所フラクタル次元

いくつかのテクスチャはある一定のパターンの繰り返しによって形成されている．このパターンの大きさが大きければそのテクスチャは単純に，小さければ複雑に見えるだろう．この複雑さの指標として修正局所フラクタル次元を採用した．修正局所フラクタル次元は局所フラクタル次元を拡張した概念で，ヒトの視覚特性（Moduration Transfer Function）を考慮した局所フラクタル次元を指す [3]．

3.4 同時生起行列から得た特徴量

ある画素 i から角度 θ の方向に距離 r だけ離れた画素点の濃淡が j である確率 $P(i,j)$ を要素とする行列を同時生起行列と呼ぶ [2]．これはテクスチャ分類に有効とされている [4]．同時生起行列を用いて 14 種の特徴量を求めた．なお同時生起行列を生成する際の階調は 256，角度は 0°，距離は修正局所フラクタル次元解析で求めたスケールとした．

3.5 差分統計量

同時生起行列を利用して，ある画素点 i から角度 θ の方向に距離 r だけ離れた画素点 j の濃淡差が k である確率から，4 種の特徴量を求めた [2]．なおこの際角度と距離は同時生起行列を用いた計算と同様にした．

[*4] C：子音（consonant），V：母音（vowel）

3.6 ランレングス行列から得た特徴量

画像内で角度 θ の方向に同じ濃淡の画素点がいくつ続いているかを要素とした行列をランレングス行列と呼ぶ [2]．これを用いて 5 種の特徴量を求めた．この際角度は 0°，濃淡の階調は 256 とした．

3.7 フーリエパワースペクトルから得た特徴量

いくつかのテクスチャは，一様な模様ではなく方向性をもっている．画像のどの方向への方向性が強いかをフーリエパワースペクトルによって求めることができる [5]．フーリエパワースペクトルの定義は式 2 で示される [2]．

$$P(\varepsilon,\eta) = |F(\varepsilon,\eta)|^2 \tag{2}$$

本研究では $P(\varepsilon,\eta)$ を極座標形式 $P(r,\theta)$ で表し，r は修正局所フラクタル次元を適応し，θ は 0° とした．

4. テクスチャ特徴量の主成分分析

テクスチャの特徴量で求めた 48 の特徴量を主成分分析した結果が図 6 である．図 6 に示された各テクスチャは一定以下の共分散の絶対値をもって分布している．つまりこれまでに示した解析によって得た特徴量は，50 のテクスチャをよく分離しているといえるだろう．

テクスチャの分散を最大とするよう設定されたこの主成分の軸をそれぞれ ζ_1，ζ_2 とする．これらをパラメータとすれば，画像ごとの差を最もよく表現できる．

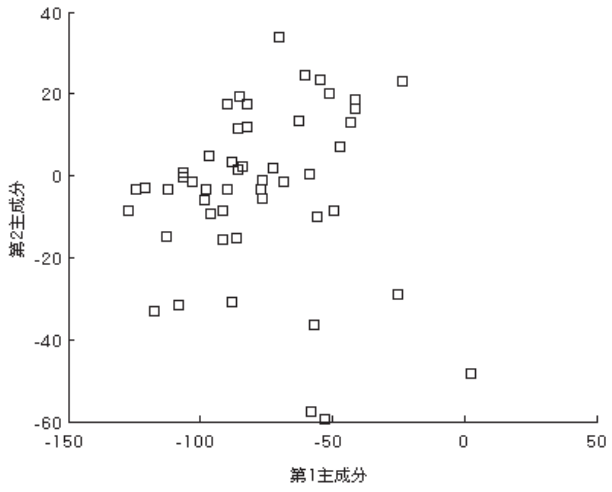

図 6 テクスチャ特徴量の主成分結果
Fig. 6 The result of PCA based on feature values of textures

5. 録音したオノマトペ解析

オノマトペが音としてどのような情報をもつのか調べるために，実際に被験者にオノマトペを発声させ，それを録音して採集した．被験者には第 2 項の実験 B で用いた 102 のオノマトペをランダム順に，マイクから 10cm の距離か

ら発声させた．

採集した音声データを音素ごとにわけ，それぞれの基本周波数 f_0，フォルマント周波数 F_1~$_3$，メル周波数ケプストラム係数（MFCC）1〜12 次を求めた．ここで得られた任意の音素の種におけるこれらの合計 16 次元の特徴量を $\boldsymbol{FV}$ とした．行列 $\boldsymbol{FV}$ は式 3 のような構成である．

$$\boldsymbol{FV} = \begin{pmatrix} /\mathrm{a}/\text{の } f_0 & /\mathrm{i}/\text{の } f_0 & \cdots & /\mathrm{p}/\text{の } f_0 \\ /\mathrm{a}/\text{の } F_1 & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ /\mathrm{a}/\text{の} & & & /\mathrm{p}/\text{の} \\ \text{MFCC12 次} & \cdots & \cdots & \text{MFCC12 次} \end{pmatrix} \quad (3)$$

5.1 基本周波数

声帯振動が生じる時間間隔の逆数を基本周波数という．これは一般にピッチに相当する物理量である [6]．対数スペクトル [*5] からケプストラム法によって導ける．

フーリエ変換によって信号の周波数成分と振幅成分を抽出したものを振幅スペクトルと呼び，解くに縦軸を対数スケールで表現したスペクトルを対数スペクトルと呼ぶ [7]．

対数スペクトルを再度フーリエ変換すると，音声の声道成分と音源特性に分離できる．この音源特性が基本周波数を表す．この処理は結局周波数領域から時間領域への逆変換であるが，位相の情報が欠落しているので元の信号とは一致しない．この手法をケプストラム法という．

5.2 フォルマント周波数

振幅スペクトルから線形予測法（LPC）を利用してスペクトル包絡を求め，そのピークを示す周波数がフォルマント周波数である．これは一般に母音の音素を識別するのに有効な特徴量とされている．本研究では LPC の次数を 64 とし，周波数が小さい順に第 1〜3 フォルマントまで求めた．

5.3 メル周波数ケプストラム係数

ヒトの音の高さの認知は，解析的に求められた周波数の変化率とは一致しない．そこで用いられるのがメル尺度と呼ばれる感覚尺度である．メル尺度においては，ヒトが 1000Hz，40db SPL の音を聴いたときに感じる音の高さを 1000mel とし，その 2 倍の高さに聴こえる音を 2000mel，1/2 の高さに聴こえる音を 500mel とする．一般に近似式 4 で表される．ただし m はメル尺度値，f は周波数を指す．

$$m = 1000 \log_2 \left(1 + \frac{f}{1000}\right) \quad (4)$$

音源をメル尺度で等間隔となる 20 の周波数領域に分割し，それぞれの振幅スペクトルを求め，それを圧縮して離散コサイン変換すると 20 次元のメル周波数ケプストラム係数（MFCC）が得られる．そのうち低次成分 12 次元を特徴量として採用した．

[*5] スペクトルとスペクトラムは同義である．英語ではスペクトルという言葉はないが，日本語においてはどちらの表現も用いられる．

6. テクスチャの表現に重要な音響特徴量

6.1 方法

ヒトが発するオノマトペのどのような音しての情報が，対象のテクスチャを区別しているのかを調べるために，距離分析を用いて音響特徴量の評価を行った．

まず距離分析を行う前に，第 2 節で得たそれぞれのテクスチャの発現順 i 番目の音素の発現確率 $\boldsymbol{P_i}$ を持ちいて，第 5 節で求めた音響特徴量を重み付けした．

$$\boldsymbol{WFV_i} = \boldsymbol{P_i FV}^{\mathrm{T}} \quad (5)$$

ζ_1 と ζ_2 からなるテクスチャ特徴量の部分空間における n 番目のテクスチャのプロット（図 6 上の座標）を $P_{n,img}(x_{img}, y_{img})$，任意の 2 つの音響特徴次元の軸に射影された座標を $P_{n,snd}(x_{snd}, y_{snd})$ としたとき，それらのユークリッド距離 d_n を式 6 で求めた．

$$d_n = \sqrt{(x_{img} - x_{snd})^2 + (y_{img} - y_{snd})^2} \quad (6)$$

このときテクスチャの第 1，2 主成分と音響特徴量は平均を同じにし，分散で正規化して行った．これは変域や分散の違いによって，距離が過大あるいは過小に算出されるのを防ぐためである．

任意の音響特徴量の組み合わせ (j, k) においてここで得られた距離 d_n の合計が最小であるとき，その音響特徴量の組は最もテクスチャ主成分と類似度が高いといえる．この類似度の評価値 $E_{j,k}$ を式 7 で定義した．

$$E_{j,k} = \frac{1}{\sum d_n} \quad (7)$$

6.2 結果

評価値 $E_{j,k}$ が最も大きかった音響特徴量を，音素の発現順に表したのが表 1 である．

表 1 評価値 $E_{j,k}$ が最大となる音響特徴量の組み合わせ

音素発現順	男声	女声
第 1 子音	f_0，MFCC 6 次	MFCC 9 次，11 次
第 1 母音	MFCC 1 次，12 次	MFCC 8 次，11 次
第 2 子音	MFCC 8 次，12 次	MFCC 8 次，12 次
第 2 母音	F_1，MFCC 9 次	MFCC 8 次，f_0

第 2 子音のみ男声と女声で結果が一致した．

6.3 考察

テクスチャを表現するのに用いるオノマトペは男女共通であると考えられるため，1 つのオノマトペから感じる印

象は性別に依らない．よって性別によって結果が異なる第1子音，第1，2母音の音響特徴量はテクスチャの表現において大きな影響をもたない．つまり結果が男女で共通した第2子音がテクスチャを表現するのに大きな役割を果たしていると考えられる．結果から，ヒトがテクスチャを表現するオノマトペを聴いたとき，その第2子音がどのようなMFCC8次，12次の値を示すかがその印象に影響しているとわかった．

様々な日本語のオノマトペにおいて母音は5種しか使われない背景から，本研究に着手する以前より，オノマトペに含まれる母音はテクスチャの表現において大きな役割を持たず，子音が重要性を持つだろうという予想があった．しかし第1子音が母音同様，重要な役割を果たさないという結果が今回確認された．つまりテクスチャを表現する音としては，同じ音響特徴量の値を持つ音源であっても，その発現する順序や前後の音が印象に大きな影響を与えていると言えるだろう．

6.3.1 音響特徴量を制御するソフトウェア開発

本研究で得られた結果は，任意の音源に対しMFCC8次，12次をパラメータとして変化させれば，その音源にオノマトペとしての印象を付加できることを意味している．そこでMFCC8次，12次を変化させるソフトウェアを開発した．

まず前提としてMFCCの解析方法からわかるように，MFCCの値だけで元の音声信号を復元するのは不可能である．これはケプストラムの解析の過程で位相の情報が失われるためである．よって本研究ではMFCCから入力音源を変化させる方法としてフィルタリングを用いた．MFCC8次，12次を操作して音源に効果を与えるフィルタを作成するため，MFCCの解析方法を逆からたどり，これらが示す周波数特性について調べた．MFCC8次，あるいは12次の値が1の音源とはどのような周波数構成なのかを調べれば，それからフィルタが作成できる．MFCCの解析方法に倣って，まずMFCC8次，もしくは12次の値が1でその他が0のデータを逆離散コサイン変換した．

さらにこのデータを振幅スペクトルに変換した結果が図7である．ここで破線はMFCC8次もしくは12次が1でそれ他が0のデータの理想応答，実線が設計したフィルタである．本研究では最小二乗線形位相FIRフィルタを用いてフィルタを設計した．低い周波数においてのフィルタは理想応答の性質を満足させない結果となった．MFCC8次，12次が示す周波数特性は周波数が大きくなるにつれ変化が小さくなり，同じ次数のフィルタ設計では双方に対応できないためだと考えられる．

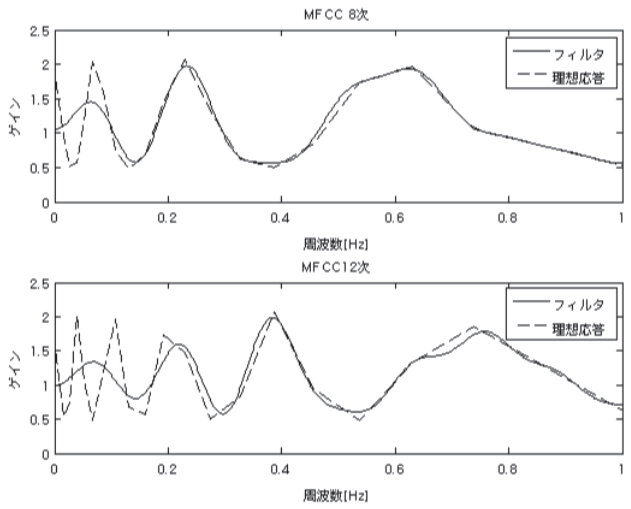

図 7 MFCC8次，または12次の特性を表すフィルタ
Fig. 7 Filters that represent the characteristics of 8th or 12th of MFCC

7. まとめ

7.1 結論

本研究では，ヒトがテクスチャをみたときそれを表現するオノマトペの分析により，そのオノマトペのどのような成分がテクスチャを表現しているのかを明らかにした．

具体的にはテクスチャを表現するにはオノマトペの第2子音のMFCC8次，12次が最も大きい役割を担うのがわかった．

MFCCの13〜20次元を削除し，1〜12次元だけで復元率の高い周波数構成を得る方法は一般的にも知られているが，本研究のように，MFCCのいずれかの次元単体に関して意味を見いだした例は少ない．本研究がMFCCの1次元ごとの意味を見いだす研究のきっかけとなるのを期待する．

7.2 検討課題

7.2.1 効果を付与する入力音源の問題

本研究で作成したのは，エフェクタの1つであるフィルタである．フィルタにはそれを通過させるための入力となる音源が必要であり，またそれがどのような音源であるべきかについては今後検討すべきである．基本となる入力音源をつくる方法として，「平均的な音素」の作成を提案する．ここでの「平均的な音素」とはどの音素にも聴こえ，またどの音素とも言い切れない曖昧な音素である．子音と母音でそれぞれ作成する必要がある．

また音素の基本の音源を作成できたなら，それを並べてオノマトペの基本の音源を作成する必要がある．音素によってはその直前，直後に無音区間が現れる特徴があり，それを忠実に再現すればオノマトペの基本音源ができる．その後オノマトペの基本音源の第2子音に，テクスチャの主成分軸と対応させたMFCC8次，12次の値から効果を与えれば，直感的にもテクスチャの印象と一致する音源を作成できるだろう．

7.2.2 音声特徴量の次元数の問題

本研究では既存の解析方法を参照し，音素について調べて利用しているが，これらの特徴量がその媒体のすべての情報を表現できているとは限らない．テクスチャの特徴解析は主成分分析の結果からわかるように，一定以下の共分散の絶対値をもって分布していることから，本研究において充分な性能を発揮していると言える．しかし音声の特徴解析において抽出できた特徴量は 16 とテクスチャ解析に比べ少なく，それらが音声の情報をすべて表現しているとは言えない．新たな音声の解析方法が開発され本研究に応用されればさらなる精度の向上が期待できるだろう．

7.2.3 フィルタの精度の問題

本研究で設計したフィルタは特に低周波数域における精度がよくない．フィルタの設計において次数の設定はその精度に大きく影響するが，MFCC8 次，12 次においては周波数帯域によって最適なフォルタの次数が異なるためいずれかの帯域における精度が満足されない．任意の周波数帯域における最適な次数のフォルタを組み合わせて設計する方法が考えられれば，設計したフィルタを本研究で求めた理想応答に近づけられるだろう．

謝辞 本研究の一部は，同志社大学理工学研究所研究助成金および JSPS 科研費 25350032 の助成を受けました．ここに記して謝意を表します．また本研究にあたって実験に参加してくださった被験者の方々に感謝いたします．

参考文献

[1] 山口仲美:暮らしのことば擬音・擬態語辞典, 講談社 (2003).
[2] 高木幹雄，下田陽久:新編画像解析ハンドブック，東京大学出版会 (2004).
[3] 山下博，中元淳:三菱重工技法 33 巻 No.6 フラクタルを応用した印刷濃度むらの定量評価法の研究，三菱重工 (2003.11).
[4] 石川達也，下野哲雄，北島秀夫，黒部貞一:同時生起行列を用いた画像分割，北海道大学工学部研究報告 105, p87-92(1981.7).
[5] 田中敏幸，村瀬曜子，上田知郎:テクスチャー特徴による肉腫の照合，日本医用画像工学会 Vol.19 No.1(2001.1).
[6] 森勢将雅:電気情報通信学会知識ベース知識の森 2 群 9 編音楽情報処理 2 章技術・アプリケーション 2-2 基本周波数推定（音声研究に関する視点から），http://ieice-hbkb.org/portal/doc_index.html，参照日 (2014.11.5).
[7] 宮澤幸希:Miyazawa's Pukiwiki 公開版”，http://shower.human.waseda.ac.jp/~m-kouki/pukiwiki_public/index.php?Index，参照日 (2014.11.1).
[8] 嵯峨山茂樹:Lectures 3. 線形予測分析 [全極音声モデル化]，OCW UTokyo OpenCourseWare，http://ocw.u-tokyo.ac.jp/lecture?id=11270&r=1280243564，参照日 (2014.11.6).
[9] 赤木正人:一般社団法人日本音響学会 Q and A(030)，http://www.asj.gr.jp/qanda/answer/30.html，参照日 (2014.11.6).
[10] 板橋秀一:音声工学，森北出版株式会社 (2005).