音声が感情を含むことによる声質変化の解析

森山　剛[*1]　　茂　美穂[*2]

Statistical analysis of Mel-Frequency Cepstrum Coefficients on emotional speech

Tsuyoshi Moriyama[*1] and Miho Shigeru[*2]

Voice timbre is often parameterized and characterized by analyzing Mel-Frequency Cepstrum Coefficients (MFCC). Human voice changes its timbre when conveying paralinguistic information such as emotions. This paper shows the result of principal component analysis (PCA) on a set of temporal sequences of MFCC that are computed from utterances that a single male speaker uttered a word /nama/ with each of 47 emotions. The scatter plot of emotions over the principal component axes indicated that the first principal component (with over 30% of contribution ratio) was the axis for the breathiness of voice, where emotions such as “fear”, “longing”, and “haste” had positive and large loadings, whereas those such as “disgust”, “hate”, and “indignation” negative and large loadings. The result also showed that whether the phonemes of interest took the role of accent nuclei affected the variation of the voice timbre due to the emotions contained.

１．はじめに

音声とは，人間がコミュニケーションのために人間の音声器官を使って発する音のことである．音声により表現され，伝達される情報は，言語的情報，パラ言語的情報，非言語的情報の 3 種類に大別される [1)].

言語的情報とは，辞書，統語，意味，談話のレベルで，文字言語に直接含まれるものを指す．パラ言語的情報とは，直接的に文字表現に含まれないが，音声言語の韻律的特徴によって話者の発話意図や感情状態などが表現されるもので，話者が意識的に，あるいは無意識的にその表現を制御できるものを指す．そして非言語的情報とは，発話内容に関係せず，話者が意識的に制御し得ない個人的特徴や身体状態などの情報を指す．このように音声には複雑な情報が含まれていることから人間のコミュニケーション手段としても特に重要なものと考えられ，工学，心理学，生理学，言語学といった様々な学問分野において研究が行なわれてきた [2)].

音声と心理状態の関連にはじめに注目したのは 1872 年の Darwin の研究と言われており，人間の本能や進化論との関係が議論されている [3)]．例えば，危機から脱出するために，興奮による筋肉の緊張で横隔膜が強く押し上げられ，従って強い呼気が発生する．これにより，強い声帯振動が誘発される，というような仮説が紹介されている．このような生理的な考察から，音声の生成モデルを仮定し，感情と音声の抑揚を含む物理的特徴との関係について，研究がおこなわれてきた [4)].

従来の感情音声に対する研究の多くは，音声の韻律的特徴を調査したものであった．音声の韻律的特徴とは，声の高・低，強・弱，リズム・テンポを指す．韻律的特徴は，発話意図などのパラ言語情報や，非言語情報を伝達する担い手と言われている．河津ら [5)]は感情音声に含まれる感情の程度と基本周波数パターンとの関係とをその生成過程モデルに基づいて分析した．平賀ら [6)]はピッチ周波数・振幅の変化パターンの検討を中心に分析した．これらの研究では共通に，感情が音声の韻律的特徴によって優位に伝達されるという一方で，「怒り」と「喜び」のように，心理としては全く異なるものであるにも関

[*1] 東京工芸大学工学部メディア画像学科助教
[*2] 富士ソフト株式会社
2011 年 9 月 20 日 受理

わらず，韻律的特徴は類似しており，その声質によってのみ区別され得るものが存在することも報告されている [7)8)].

本研究では，感情を含むことによって，音声の声質がどのように変化するかを明らかにするために，音声中のアクセントのある音節（アクセント核）とそうでない音節とに分けて，それぞれ，複数の感情を含んだ音声の集合について，声質パラメータの主成分分析を行う．主成分に対する解釈を行うことにより，感情を含むことによって変化する主たる声質成分を明らかにする．

2．感情を含むことによる声質変化

2.1 感情に特徴的な声質

従来の研究では，音声が感情を含んだ際，基本周波数，パワー，持続時間などの韻律的特徴を分析したものが多く，またケプストラムなどのスペクトル情報に基づいた分節的特徴を利用したものも存在する．しかし上述のように，感情には，「怒り」や「喜び」のように韻律のみでは識別が困難であると報告されているものがある．このような感情は抑揚が似ていても声質が異なる．例えば「怒り」では，力強く凄みのある声質が特徴的であり，また「喜び」では，明るく甲高い声質が特徴的である．また「呆れ」や「嘆き」を含んだ音声では，持続時間が長くなる点は共通だが，声質が末尾に向かって徐々に変化する様子が異なる．このように，感情に特徴的な声質の変化とは，静的な周波数構造の変化と，それが時間的に変動する動的な変化の組合せであると考えられる．

ここでは，動的な変化のうち，アクセント核の位置に起因する変動に注目して，音声が感情を含む際の声質の変化に関する分析を行うこととする．また，静的な周波数構造を表現するパラメータとして，次節で述べるメル周波数ケプストラム係数を用いることとする．

2.2 声質の特徴量

声質は，メル周波数ケプストラム係数(Mel-Frequency Cepstrum Coefficients: MFCC)でパラメータ化されることが多い．MFCC は，音声波形のスペクトルを人間の聴覚に近い周波数間隔に切り分けてスペクトル包絡情報を表すための特徴量である．人の聴覚は，低い周波数では細かく，高い周波数では粗い周波数分解能をもつことが知られている [9)-11)]．これはメル（mel）尺度と呼ばれ，対数に近い非線形の特性を示す．メル尺度 f_{mel} は(1)式で表わされる．

$$f_{mel} = 2595\log_{10}\left(1+\frac{f}{700}\right) \tag{1}$$

MFCC の計算方法には様々な方法があるが，ここではフィルタバンク分析を利用した手法を用いる．この手法では，まず音声波形に対して高速フーリエ変換（Fast Fourier Transform : FFT）処理を施す．次に，メル尺度に基づいた L 個の三角窓による帯域フィルタを周波数軸に配置し，FFT スペクトルと掛け合わせる．最後に(2)式に示すように，フィルタバンク分析により得られた L 個の帯域におけるパワーに対して離散コサイン変換（DCT）を行うことで MFCC を得ることができる．

$$c_i = \sqrt{\frac{2}{N}}\sum_{l=1}^{L} A_l \cos\left[\left(l-\frac{1}{2}\right)\frac{i\pi}{L}\right] \tag{2}$$

N はフレーム長，A_l は対数フィルタバンクの振幅，i は MFCC の次数（$i=1,\ldots,p$）を表す．

3．声質特徴量の解析

3.1 声質行列の生成

p 次元の MFCC を，K フレームに渡って算出する．この $p\times K$ 個の係数すべてを独立変数とする声質変数ベクトルを(3)式のように生成する．

$$\boldsymbol{c}_j = \{c_{1,1},\ldots,c_{1,p},c_{2,1},\ldots,c_{2,p},\ldots,c_{K,1},\ldots,c_{K,p}\}_j \tag{3}$$

これを J 個の音声について求め，(4)式のように，$\boldsymbol{c}_j^T$（T は転置を表す）を並べてできる行列 $\boldsymbol{Q}$ を声質行列と呼ぶこととする．

$$\boldsymbol{Q} = \left[\boldsymbol{c}_1^T \middle| \boldsymbol{c}_2^T \middle| \cdots \middle| \boldsymbol{c}_J^T\right] \tag{4}$$

3.2 声質特徴量の主成分分析

主成分分析とは，ある変数群に対してその変数相互の関係から新しい概念のファクタを導き出す統計的手法であり，このファクタによって変数の類似

性やポジショニングを明らかにするものである．声質行列 $\boldsymbol{Q}$ に対して主成分分析を行うことで，次式が成立する．

$$\begin{aligned} c_{1,1} &= \bar{c}_{1,1} + a_{1,11}x_1 + a_{1,12}x_2 + \cdots + a_{1,1m}x_m \\ c_{1,2} &= \bar{c}_{1,2} + a_{1,21}x_1 + a_{1,22}x_2 + \cdots + a_{1,2m}x_m \\ &\vdots \qquad\qquad \vdots \\ c_{1,p} &= \bar{c}_{1,p} + a_{1,p1}x_1 + a_{1,p2}x_2 + \cdots + a_{1,pm}x_m \end{aligned} \tag{5}$$

(5)式において $x_1, x_2, \ldots, x_m$ は固有ベクトル，$a_{1,11}, a_{1,21}, \ldots, a_{1,pm}$ は主成分得点，$c_{1,1}, c_{1,2}, \ldots, c_{1,p}$ は基準化された変数の値，$\bar{c}_{1,1}, \bar{c}_{1,2}, \ldots, \bar{c}_{1,p}$ は変数のサンプル間の平均値，m は $p \times K$ 個の全主成分のうち使用する主成分の数である．主成分得点は，主成分空間における軸を表す固有ベクトルの各要素に各主成分の固有値を乗算したものであり，各変数の各主成分への重みを決定している．固有値の大きい主成分から順に第 1 主成分，第 2 主成分，…，第 m 主成分と呼ぶ．

3.2.1　主成分数の決定法

主成分得点の分散である固有値が大きいほど，その主成分の説明力が高くなる．第 k 主成分の固有値 λ_k が固有値の合計に占める割合を寄与率といい，次式で表わされる．

$$r_k = \frac{\lambda_k}{\sum_{l=1}^{p \times K} \lambda_l} \tag{6}$$

また，第 k 主成分までの固有値の合計が全固有値の合計に占める割合を累積寄与率といい，抽出する主成分を選択する尺度として用いる．一般に，

- 累積寄与率が 70%を超える主成分まで抽出する．
- 固有値が 1.0 を超える主成分までを抽出する．

などが抽出する主成分の数 m の決定方法として使われている．本研究では 70%を超える主成分まで抽出する．

4．声質変化の解析実験

様々な感情を含んだ音声を用意し，それらすべての音声の MFCC 時系列を集約して声質行列の形に変換し，これを主成分分析することにより，感情を含むことによって変動する主要な声質成分を，主成分として抽出した．主成分分析の結果に関して，アクセント核に対するものと非アクセント核に対するものとの比較を行った．

4.1　実験に用いた音声データ

話者は，男性話者 1 名とした．

発話内容は，感情情報に関する従来の研究では，単文を用いたものがあるが，本研究では文脈の影響を排除するために，文節レベルのアクセント句とした．感情音声データの発話内容として，ある特定の感情に属するような内容ではないこと，話者が発話すること，感情を込めることが容易であること，の 3 つが満たすべき条件として挙げられる．ここでは，同じ母音が，アクセント核とそうでない位置とに含まれるものとして，「なま/nama/」という 2 モーラ頭高型の語を選んだ．

話者の込める感情については，バリエーション豊かな感情を含む必要がある．また，話者にどのような感情を込めるか何も手掛かりを与えないと，MFCC の分布が偏ってしまう可能性がある．そのため，話者の感情表現の手掛かりとして表 1 に示す 47

表 1　日常の感情を表現する言葉

1.	平静	17.	気の毒な	33.	媚び
2.	怒り	18.	寛容	34.	満足
3.	喜び	19.	ほくそ笑む	35.	退屈
4.	嫌悪	20.	失望	36.	苦しい
5.	侮り	21.	叱責	37.	期待
6.	おかしい	22.	悲しい	38.	幸福
7.	心配	23.	恐れ	39.	好き
8.	優しい	24.	憎い	40.	嫌い
9.	安堵	25.	軽蔑	41.	いや
10.	憤慨	26.	嬉しい	42.	落胆
11.	羞恥	27.	皮肉	43.	非難
12.	穏やか	28.	無関心	44.	不安
13.	憧れ	29.	賞賛	45.	驚き
14.	苛立ち	30.	誇り	46.	慌て
15.	不平	31.	愛	47.	呆れ
16.	切望	32.	嘆き		

表 2　音声の分析条件

窓関数	ハミング窓
フレーム長	256 サンプル（$N = 256$，20ms に相当）
フレーム間隔	128 サンプル（10ms に相当）
MFCC 次数	13（$p = 13$）

種類の感情語を教示した（(4)式で $J = 47$）．

音声波形離散化の条件としては，サンプリング周波数 16kHz，16bit 線形量子化で A-D 変換し，ファイルに保存した．

音声の切り出しは，収録音声/nama/のアクセント核/na/の/a/を以下/na/，同様に/nama/の非アクセント核/ma/の/a/を以下/ma/と表記する．そして/na/と/ma/の 2 つの区間を，47 感情の音声すべてについて切り出して分析に用いた．

4.2　MFCC の分析条件

MFCC の分析条件を表 2 のようにした．また，すべての音声の長さを 20 フレーム分（K=20; 160ms に相当）に正規化し，音声を 13 次の MFCC，20 フレーム分からなる 260 次元のベクトルで表し，これを転置し，47 音声分（47 列）を合わせて声質行列 $\boldsymbol{Q}$ を生成した．そしてこの声質行列 $\boldsymbol{Q}$ について，主成分分析を行った．

4.3　実験結果

4.3.1　アクセント核に関する結果

アクセント核/na/の分析を行った結果，得られた各主成分の寄与率及び累積寄与率を表 3 に示す．第 1 主成分と第 2 主成分，第 3 主成分がそれぞれ 31%，28%，12%と大きな寄与率を示した．また，累積寄

表 3　アクセント核/na/に関する主成分分析結果

	寄与率[%]	累積寄与率[%]
第 1 主成分	31	31
第 2 主成分	28	59
第 3 主成分	12	71
第 4 主成分	5	76
第 5 主成分	5	81
…	…	…

表 4　マーカと対応する感情語

平静	△	安堵	○	気の毒な	×	軽蔑	*	媚び	□	いや	◇
怒り	△	憤慨	○	寛容	×	嬉しい	*	満足	□	落胆	◇
喜び	△	羞恥	○	ほくそえむ	+	皮肉	*	退屈	□	非難	▽
嫌悪	△	穏やか	○	失望	+	無関心	*	苦しい	□	不安	▽
侮り	△	憧れ	×	叱責	+	賞賛	*	期待	◇	驚き	▽
おかしい	△	苛立ち	×	悲しい	+	誇り	*	幸福	◇	慌て	▽
心配	○	不平	×	恐れ	+	愛	□	好き	◇	あきれ	▽
優しい	○	切望	×	憎い	+	嘆き	□	嫌い	◇		

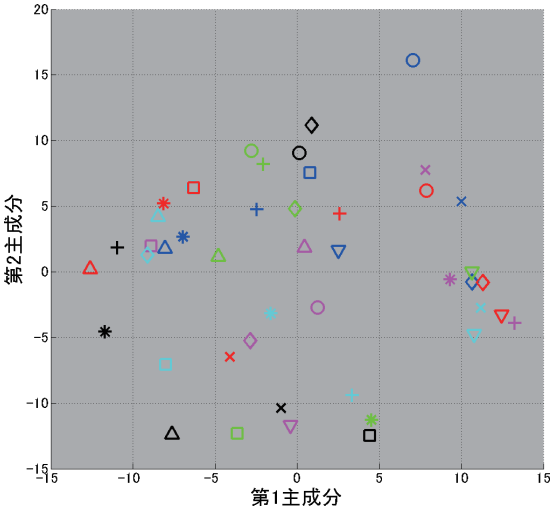


図 1　第 1—第 2 主成分平面への感情の布置

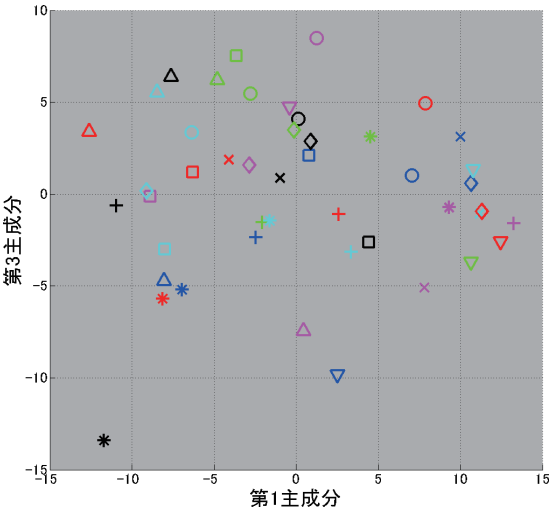


図 2　第 1—第 3 主成分平面への感情の布置

与率は第 5 主成分までで 80%を超える結果となった．

図 1 に横軸を第 1 主成分，縦軸を第 2 主成分とした感情ごとの声質ベクトル（主成分得点ベクトル）の布置図を示す．図 2 に横軸を第 1 主成分，縦軸を第3主成分とした感情ごとの声質ベクトルの布置図を示す．図 1 及び 2 におけるマーカとそれらに対応する感情語を表 4 に示す．両布置図において，「恐れ」「慌て」「切望」は近傍にプロットされ，またその対極に「喜び」「憎い」「誇り」がプロットされた．前者の音声は息混ぜでスピード感をもって発話されているのに対して，後者の音声は逆に息をあまり使わずに発話される特徴を有する．従って，第 1 主成分は「息混ぜ」に関する軸と解釈できると考えられる．

4.3.2 非アクセント核に関する結果

非アクセント核/ma/の分析を行った結果，得られた各主成分の寄与率及び累積寄与率を表 5 に示す．第 1 主成分と第 2 主成分，第 3 主成分がそれぞれ 30%，24%，18%と大きな寄与率を示した．また，累積寄与率は，アクセント核に関する結果と同様に，第 5 主成分までで 80%を超える結果となった．

図 3 に横軸を第 1 主成分，縦軸を第 2 主成分とした感情ごとの声質ベクトルの布置図を示す．図 4 に横軸を第 1 主成分，縦軸を第 3 主成分とした感情ごとの声質ベクトルの布置図を示す．両布置図上において，「愛」「安堵」「落胆」といった感情が，第 1 主成分軸の正に大きな得点を有し，「苦しい」「悲しい」「誇り」といった感情が，その対極において負の大きな得点を有している．前者がゆったりと発話されるのに対し，後者は切迫感をもって発話される感情が集中している．そこで，非アクセント核における最も主要な声質の変化は，「穏やかさ」軸であると解釈できると考えられる．

4.3.3 声質の主成分分析結果に関する考察

13 次元の 20 フレームにわたる時間軌跡 260 次元で表わされていた声質を，主成分分析を行う事により 5 次元程度の低次元で表わせることがわかった．

また，第 1 主成分に関して，布置図上で両端に分布している感情を調査した結果，アクセント核/na/においては「息混ぜ」による声質変化が最も主要な

表 5 非アクセント核/ma/に関する主成分分析結果

	寄与率[%]	累積寄与率[%]
第 1 主成分	30	30
第 2 主成分	24	54
第 3 主成分	18	72
第 4 主成分	6	78
第 5 主成分	3	81
…	…	…

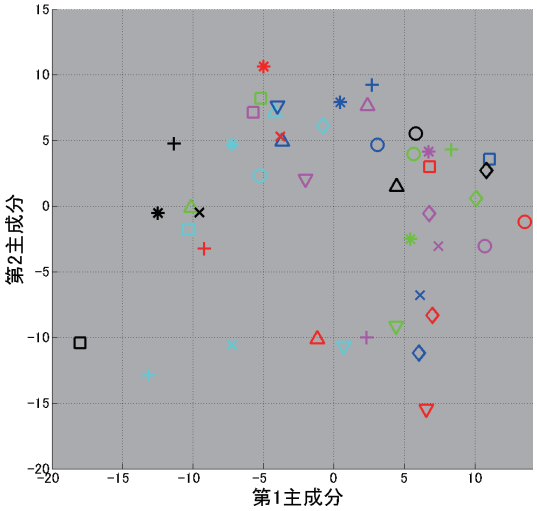


図 3 第 1―第 2 主成分平面への感情の布置

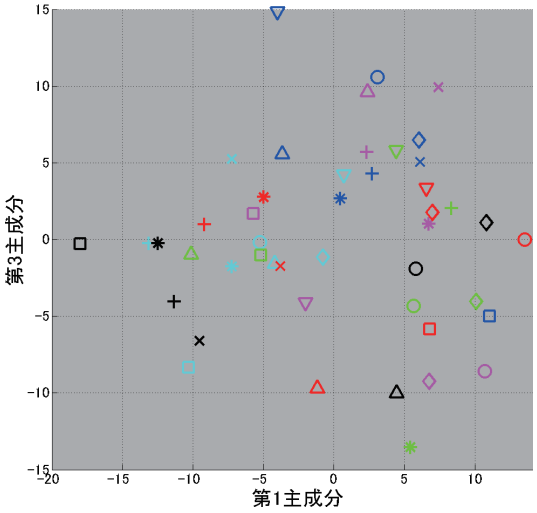
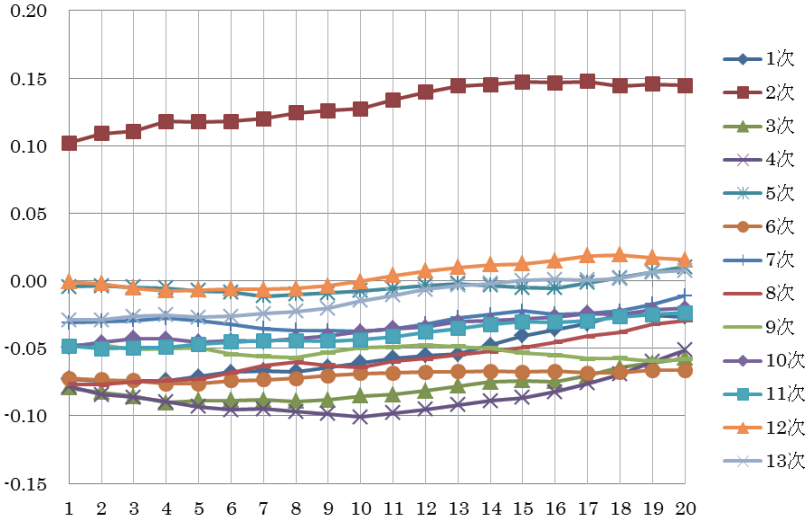
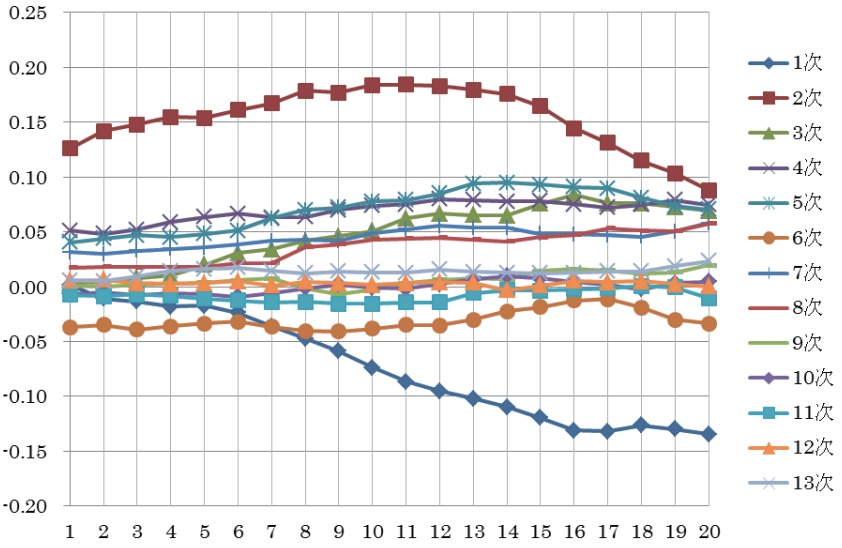


図 4 第 1―第 3 主成分平面への感情の布置



(a)　アクセント核/na/について



(b)　非アクセント核/ma/について

図 5　第 1 主成分に対する主成分負荷量

ものであり，一方で，非アクセント核/ma/においては「穏やかさ」による声質変化が最も主要なものであることがわかった．

また，第 1 主成分への主成分負荷量 260 次元を 13 次の MFCC それぞれを 20 フレームに分解したものを図 5（アクセント核/na/に関するものを(a)，非アクセント核/ma/に関するものを(b)）に示す．図 5(a) より，アクセント核/na/では，上位の係数が 2 次，12 次，13 次，7 次といったように周波数帯域の全体に渡っていることがわかる．これに対して非アクセント核/ma/の図 5(b)では，上位の係数が 1 次～5 次の低域に集中している（1 次は負であるが絶対値が大きい）ことがわかる．すなわち，音声に含まれる感情が異なると，非アクセント核における声質の，

特に低域の変動が最も分散に寄与することが明らかとなった．また，アクセント核，非アクセント核双方で 2 次の係数が大きな負荷量を持つことから，この帯域が声質の変化を最も受けると言える．

５．まとめ

音声の声質が，感情を含むことによってどのように変化するかを，様々な感情を含んだ音声のメル周波数ケプストラム係数 MFCC の時間変動に対して主成分分析を行うことにより（限られた音声条件ではあるが）明らかにした．

実験の結果，音節がアクセント核であるか否かによって，感情を含むことによる声質の変化に関して，主要な成分が異なることが示された．具体的には，実験で用いた言葉/nama/において，アクセント核/na/については，第 1 主成分の両端に「恐れ」「慌て」「切望」が正方向，「喜び」「憎い」「誇り」が負方向に布置され，非アクセント核/ma/については，「愛」「安堵」「落胆」が正方向，「苦しい」「悲しい」「誇り」が負方向に布置された．これらの音声の特徴から，第 1 主成分は，アクセント核では「息混ぜ」軸，非アクセント核では「穏やかさ」軸と解釈されると考えられる．また，第 1 主成分に対する負荷量から，非アクセント核では低域，アクセント核か否かに関わらず 2 次の係数の表す帯域が，感情による声質の変化を最も表すことがわかった．

本実験により，13 次元の 20 フレームにわたる時間軌跡260次元で表わされていた声質パラメータが，5 次元程度の主成分によって表現できることから，低次元のパラメータから，含まれている感情を推定したり，これらのパラメータを操作することにより，感情を含んだ音声を合成したりすることも可能となると考えられる [12]．

今後は，モーラ数とアクセント型の他の組み合わせに関しても実験を行うと同時に，MFCC を主成分空間で変形することによって，自然な声質変化を伴う感情音声の合成を試みる．

参考文献

1) 藤崎博也，韻律研究の諸側面とその課題，日本音響学会，音講論集，287-290, 1994.
2) 鈴木朋子，音声と心理との関連についての音声物理量の解析，横浜国立大学大学院工学研究科平成 16 年度博士論文，2004.
3) チャールズ・ダーウィン，浜中浜太郎（訳），人及び動物の表情について，岩波文庫，1931.
4) 森山剛，音声に含まれる感情情報と物理的特徴量に関する研究，慶應義塾大学大学院理工学研究科，博士論文，1999.
5) 河津宏美，長島大介，大野澄雄，生成過程モデルに基づく感情表現における F_0 パターン制御規則の導出と合成音声による評価，電子情報通信学会論文誌．D，情報・システム，**89**(8), 1811-1819, 2006.
6) 平賀裕，斎藤善行，森島繁生，原島博，音声に含まれる感情情報抽出の一検討，電子情報通信学会，技術研究報告，ヒューマンコミュニケーション，**93**(439), 1-8, 1994.
7) 森山剛，斎藤英雄，小沢慎治，音声における感情表現語と感情表現パラメータの対応付け，電子情報通信学会論文誌．D-II，情報・システム，II-パターン処理，**82**(4), 703-711, 1999.
8) 森山剛，細田康弘，小沢慎治，音声における感情情報の認識・合成システム，電子情報通信学会ソサイエティ大会講演論文集，339, 1998.
9) 佐々木太郎，北村正，岩田彰，2 次元メルケプストラムを用いた不特定話者212単語認識法の検討，電子情報通信学会技術研究報告. SP，音声，**93**(266), 47-54, 1993.
10) 板橋秀一，音声工学，森北出版，2005.
11) 鹿野清宏，音声認識システム，オーム社，2006.
12) 石田堅修，森山剛，小沢慎治，自然な合成音声のための調音制御規則の検討，電子情報通信学会技術研究報告. SP，音声，639, pp.15-22, 1999.