情報処理学会東北支部研究報告
IPSJ Tohoku Branch SIG Technical Report

2016/3/7 B1-4

ケプストラム分析と主成分分析による音声の感情分析

庄子雄貴† 安藤敏彦†

仙台高等専門学校

本研究は，人と人工物との相互作用を，感情を通して高める事を目標に，人の音声の感情の特徴を明らかにする事を目的とする．本稿では，感情音声コーパスを用い，音声をケプストラム分析と主成分分析を用いて分析した結果について述べる．その結果，高次のフォルマントに関わる主成分が感情の判別に有効である事が分かった．

Analysis of Speech Emotion Using Cepstrum Analysis and PCA

Yuki Shoji† and Toshihiko Ando†

The object of our research is to make clear a characteristic of each speech emotion, for enhancement of interaction between human beings and artifacts through emotion. We show the result of the analysis of speech emotion to emotional voices in an emotional voice corpus using cepstrum analysis and principal component analysis in this paper. As a result of it, we could get the fact that some principal components are effective to speech recognition related to higher formants.

1. はじめに

1. 1 研究背景

近年，ソフトバンク社から人の感情を数値化して，次の行動を決定する人口知能を搭載する「Pepper」や，独自のコミュニケーションシナリオを用いてより人間らしい自然な会話を実現した「Robi」など人と社会的にコミュニケーションが取れる人工物が増加傾向にある．

今後予想されるロボットなどの人工物の人間社会への普及とともに人と人工物の間の社会的共存のあり方を，著者らのグループで行う人工物演劇プロジェクトで探っている．人工物演劇プロジェクトでは，人と，ロボットやエージェントなどの人工物の日常のコミュニケーションをデザインするという観点から，演劇の手法を用いて両者のコミュニケーションのあり方を探っている．このプロジェクトを通して，著者らは人工物が人の音声や動作から感情を理解し，また人工物自身がそれらによって感情を表現することができれば，人工物から人にコミュニケーションをとるようにアプローチすることや，人も感情をもつ人工物に興味を示し，人工物と人が社会的に共存しやすくなるのではないかと考えている．現在，それを実現する手段として，人が行う音声や身体動作の感情表現から感情を認識し，それに対応する人工物の応答動作を生成に取り組んでおり，著者らは音声からの感情認識を担当している．

音声感情認識では，カテゴリ化によるラベルづけと感情空間へのマッピングの２つの立場で研究が行われている[1]．例えば、前者では基本周波数や持続時間，アクセントなどの韻律的特徴を用いた分析 [2] やニューラルネットワーク，サポートベクターマシンを用いた手法が行われている．野田ら[2]は基本周波数や持続時間，アクセントなどの韻律的特徴を用いて分析を行っているが，韻律的特徴は個人差が大きく感情モデルの構築が難しいようである.また自然言語解析から感情を認識する手法もあるが，現在の自然言語解析技術は明瞭な発声で読み上げた音声や，ある程度の騒音下で発声した音声は概ね正しく認識できるようになっている．一方，後者の 2 ないし 3 次元の感情空間へのマッピングは，全ての感情をいくつかの基本因子の合成として扱う立場

† 仙台高等専門学校
National Insutute of Technology, Sendai College

である．例えば，Russell [3] 感情を"Arousal" と"Valence" の２軸で張られた空間上に円環で表現している．実際，日常の感情はいくつかの感情の合成であること，「怒り」，「喜び」など同じカテゴリに分けられても感情の程度も強弱さまざまであることから，微妙な感情の認識にはこの手法の方が有利と考えられる．

人工物演劇プロジェクトでは，音声と身体動作を共通の枠組みで感情を認識，生成することを考えているので，感情空間でのマッピングの立場をとり，感情を arousal, valence の 2 軸の空間上の点として扱う予定である．自然言語解析から感情を認識する手法では，人が「考えながら発声した言葉」や「友人同士の対話音声」についての認識性能が低いので，人工物演劇プロジェクトでは，その手法上「対話音声」の認識ができないのは都合が悪い．そのため，著者らは感情認識に，個人間で差が出ず，対話音声においても比較的認識しやすい音響特徴量であるフォルマントを用いる．ただし，現時点では，各感情の音声がどのような特徴をもつかを把握するため，カテゴリ化して，それぞれの感情の特徴を調べている．本稿でもその立場から報告する．

1．2　研究目的

本研究では，ロボットなどの人工物が人の音声感情を認識し，また，人工物の音声によって人に感情を喚起させることを目標に，そのための準備として，感情を有する音声の特徴を明らかにする事を目的とする．本稿では，あらかじめラベルづけされた感情音声コーパスを利用し，ケプストラム分析によって得られたフォルマントを比較する．さらに，主成分分析を用いて，感情を判別できるような基本指標を導く．

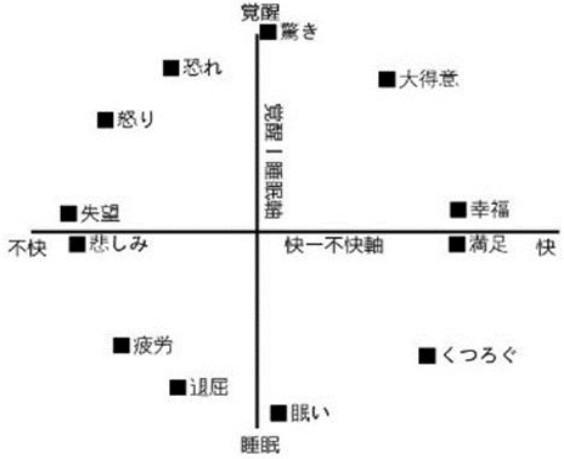

図 1．ラッセルの円環モデル

2．音声のケプストラム分析

本研究での音声から感情を認識する方法として，「怒り」「喜び」「悲しみ」「ニュートラル（平静）」の 4 つの感情の音声データを用いて，その音声データに対してまずケプストラム分析を適用し，音声の第 1 フォルマント（F1）から第 6 フォルマント（F6）とそれに対応するピーク値（AP）を抽出する．ケプストラム分析は他の音声のモデル化に基づく分析方法や，分析合成に基づく手法と比較して，計算時間が短く，正確にフォルマントが抽出できる．なぜ F6 の高次のフォルマントまで抽出するのかというとフォルマントは F1，F2 は単語，主に母音や子音などの情報があり，それ以降の高次のフォルマントには感情などの情報がある[4]と考えられているからである．

そして，そのデータ群に主成分分析を行い，次元を削減し，新たな基本指標を作り出し，グラフにプロット，感情ごとに判別関数などを用いて判別できないか試みる．

本研究では，個人間で差がでず，対話音声においても比較的認識しやすい音響特徴量であるフォルマントを用いて感情の認識を試みる．音響特徴量を分析し，グラフにプロットし，感情ごとに判別分析できるかを調べ，感情モデルを作成する．感情モデルは人工物から音声の生成を容易に行なうために，すでに表情から感情を認識する研究[3]でラッセルの円環モデル（図 1）があるので，その立場で感情モデルを作成する．

本節では，ケプストラム分析の方法と，それによって得られた結果について述べる．

2．1　研究環境及び使用音声データ

本研究では MATLAB という数値計算用ソフトウェアを使用する．MATLAB は信号処理，通信制御システム，金融工学などの分野のデータを解析するのに用いられており，音声を扱う関数が一通り揃っているのでこれを使用した．

また，本研究で使用する音声データは，国立情報学研究所が設置する音声資源コンソーシアムが提供する慶應義塾大学研究用感情音声データベース（Keio-ESD）である．音声データは 20 種類の単語に，それぞれ 47 種類の感情があるが，本研究では「怒り」「喜び」「悲しみ」「ニュートラル（平静）」の 4 つの感情を使用する．音声データの条件を以下に示す．

表 1．サンプル音声データ

話者	舞台経験のある男性（32 歳）
サンプリングレート	16kHz
フォーマット	WAV ファイル形式
量子化 bit 数	16Bit

表 2．使用音声単語

amagaeruwa	amamizu	amamizuwa	amarimonogawa	amarimono
arawani	arayuru	emoiwarenu	iwazumogana	midori
nagame	nama	nami	naname	naniyorimo
oborozukiyo	omonaga	omoumamani	warawaremono	yawarageru

2．2　ケプストラム分析の方法

人は発声するときに肺に空気を溜め，吐き出すことで声帯を振動させる．この声帯の振動ではブザーのような小さい音の振動でしかない．その振動を共振させ増幅して人は声を出している．声帯の振動をスペクトル微細構造，声道の振動をスペクトル包絡という．

ケプストラム分析とは，その 2 つの振動を分離して，スペクトル包絡を分析することである．手順としては，音声波形データをフーリエ変換（FFT）して各周波数量を得る．それで得た周波数量の対数をとり，2 つの波形を分離，さらに逆フーリエ変換をした波形データである．この工程によりスペクトル微細構造とスペクトル包絡を分離することができる．スペクトル包絡を分析する理由としては，一般的にスペクトル微細構造は個人の声の高さや有声音か無声音かの情報しか持っていないと言われていて，それに比べ，スペクトル包絡は声道の形状や周波数情報，発声した単語などの情報を持っているので音声認識の分野では重要視されている．

なぜ対数をとることにより 2 つの波形を分離できるのかというと，スペクトル微細構造を G(Ω)，スペクトル包絡を H(Ω)，この二つが畳み込まれて生成されている音声スペクトルを S(Ω)とすると，

$$|S(\Omega)| = |G(\Omega)| \cdot |H(\Omega)|$$

これの両辺の対数をとると

$$\log|S(\Omega)| = \log|G(\Omega)| + \log|H(\Omega)|$$

となる音声対数スペクトルがスペクトル構造とスペクトル包絡に分離することができる．

この分離分離した波形データに対して，逆フーリエ変換をしてスペクトル包絡のみを取り出すリフタリングという作業をし，分析することがケプストラム分析である．

本研究ではスペクトル包絡から音声のフォルマントとそのピーク値を取り出し，分析する．

ただし，ケプストラム分析を行うときは FFT をするため，その前に高域強調処理（プリエンファシス）と窓関数の適用をする．

●高域強調処理（プリエンファシス）

高域強調処理とは，フォルマント周波数が低周波数だと FFT 時に観測しにくくなってしまう可能性があるので音声データ全体の周波数を高める処理である．x(n)をサンプル数 n を変数とする音声波形，p はプリエンファシス係数（0.97）として，1 サンプル前の値と現在のサンプルの値の差分をとることによって全体の周波数を強調する．

式は以下のようになる

$$y(n) = x(n) - px(n-1)$$

母音「あ」に対して高域強調処理を適用した図 2 を示す．

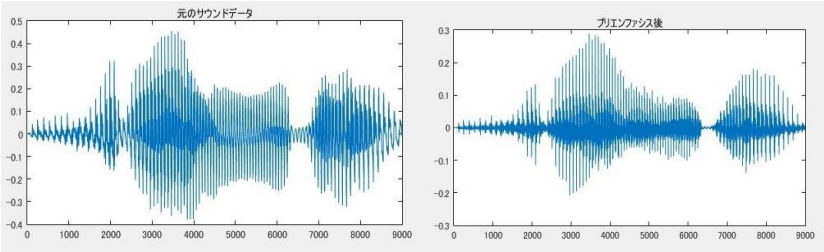

図 2．高域強調処理の前（左）と後（右）

●窓関数

FFT は周期関数に適用することが定義されている，しかし一般の音声データは周期関数になっていない．周期関数でないデータに FFT を適用するとピーク値が拡散されてしまってフォルマント周波数の正確な値が求められなくなってしまう可能性がある．よって窓関数という波形を音声データに掛けることにより音声データの両端を 0 に近似させ，拡散を抑える．図 3 に窓関数を示す．さらに，この窓関数を母音「あ」に窓関数を掛け合わせた波形を図 4，図 5 に示す．

このように音声データに高域強調処理と窓関数を適用したものにケプストラム分析をしてフォルマント周波数とピーク値を求める．母音「あ」に対してケプストラム分析を適用し，スペクトル微細構造とスペクトル包絡を分離した波形を図 6 に示す．

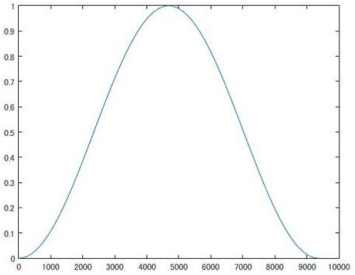

図 3. 窓関数

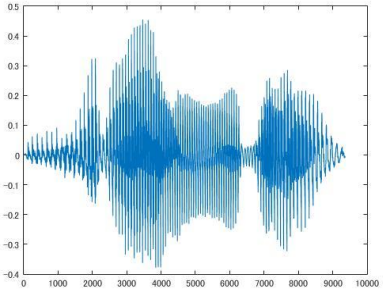

図 4（左）.「あ」波形データ

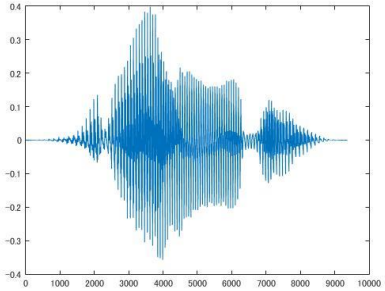

図 5（右）. 窓関数適用後「あ」波形データ

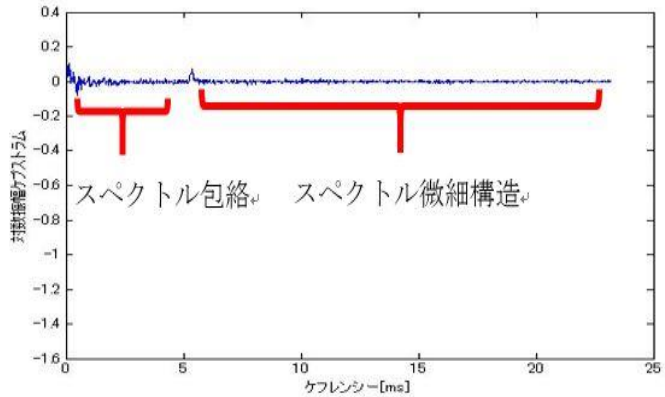

図 6. ケプストラム波形

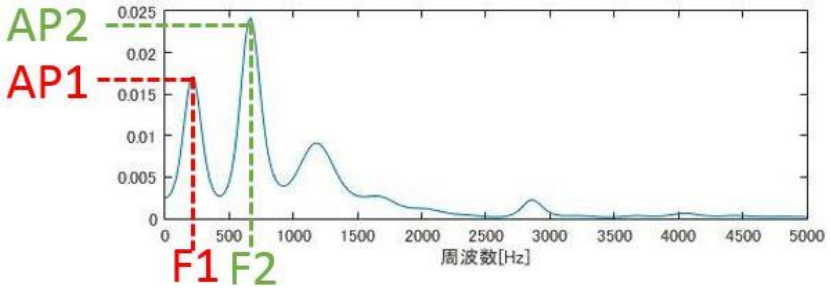

図 7. フォルマント周波数とピーク値

ここで，図 6 の縦軸は対数をとっているので逆フーリエ変換をしても元の信号にはならずにケプストラムと呼ばれ，横軸は時間軸ではなくケフレンシーと呼ばれる．図 6 の前半のスペクトル包絡の部分を取り出すためにリフタというローパスフィルタを適用する．これをリフタリングという．リフタリングをするとフォルマント周波数とそのピーク値が抽出される．それらを抽出したものを図 7 に示す．

図 7 の周波数の山なりになっている部分の周波数をフォルマント，フォルマントに対応する縦軸の数値をピーク値という．フォルマントは周波数の低い順に第 1 フォルマント，第 2 フォルマント・・となる．

このケプストラム分析を上述の「怒り」,「喜び」,「悲しみ」,「ニュートラル」の 4 種の感情音声に適用する．この分析では，第 6 フォルマントまでの周波数とそのピーク値を抽出する．4 つの感情ごとに，20 単語で得られたフォルマントの周波数とそれに対応するピーク値をグラフにプロットした結果を図 8 － 11 に示す．それぞれ，横軸が周波数，縦軸がピーク値である．

これらの図を見ると各感情のフォルマントとピーク値の散布図には赤い点線のように感情ごとに異なるパターンが観測できた．この感情ごとのパターンの違いを比較したいのだが，第 6 フォルマントとそのピーク値の成分があるということで扱う次元の量が全部で 12 次元あり，視覚的に観測するのは困難だという問題がある．

3. PCA による音声感情の分析

2 節で抽出したデータ群の次元の問題を解決するために主成分分析（PCA）を用いる．主成分分析とは，多変量データを統合し、新たな総合指標を作り出す手法で，多くの変数に重みをつけて少数の合成変数を作るのが目的である．2 節で得られた 12 次

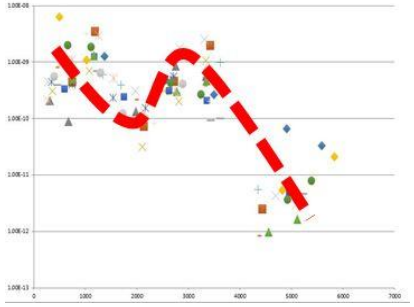

図 8（左）.「怒り」F-AP

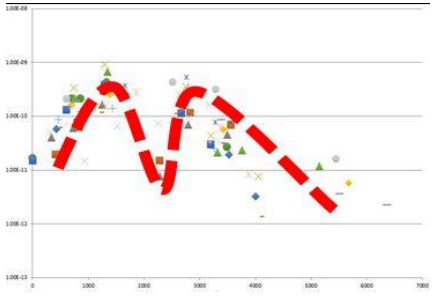

図 9（右）.「喜び」F-AP

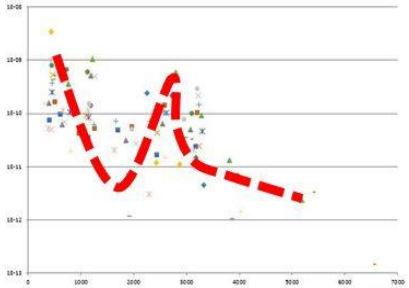

図 10（左）.「悲しみ」F-AP

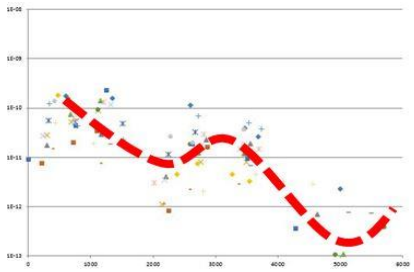

図 11（右）.「ニュートラル」F-AP

元の情報を解釈しやすいように，次元を 2，3 次元に削減していく．

ケプストラム分析で得たフォルマントとピーク値のデータを主成分分析した結果を図 12 に示す．図 12 の第 1 主成分 (PC1) が F1 から F6 までの総合的なフォルマントの高さ，第 2 主成分 (PC2) が AP1 から AP6 までの総合的なピーク値の高さを表している．また第 1 主成分と第 2 主成分の寄与率は 70%以上だったのでこの 2 つの主成分で散布図を作成した．主成分分析の結果は「怒り」と「ニュートラル」は感情がまとまっていたが，「喜び」と「悲しみ」の感情が入り乱れていたため，感情を認識する判別関数の作成は困難だと判断した．

ところが，2 種類の感情ごとに特定の主成分の組で散布図を作成してみたところ，ある特定の主成分でグループ化できる感情の組を発見した．例えば，「怒り」「喜び」は第 2 主成分 (PC2)，第 3 主成分 (PC3)で分けられ，「喜び」「悲しみ」は第 4 主成分，第 5 主成分で分別することができた．今回「ニュートラル」においては人がなにも感じてない場合，無感情と考え，散布図作成から除外した．

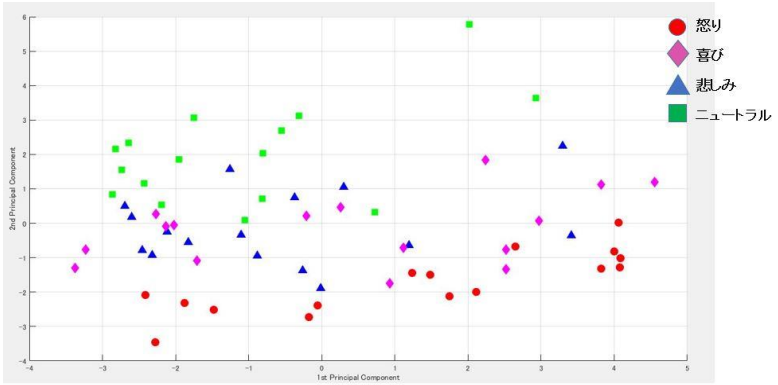

図 12．F-AP 主成分分析結果（PC1- PC2）

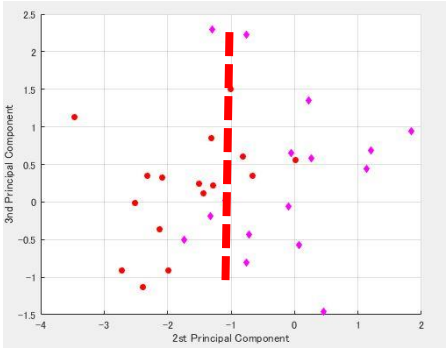

図 13（左）.「怒り」「喜び」散布図 (PC2 – PC3)

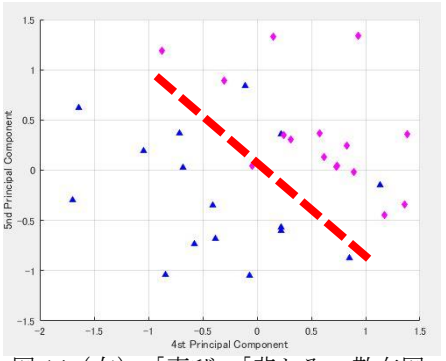

図 14（右）.「喜び」「悲しみ」散布図 (PC4 – PC5)

上図を見ると図 13 では第 2 主成分で感情が分けられていて，図 14 では第 4 主成分 (PC4)，第 5 主成分 (PC5) の線で感情が分けられている．これらの主成分で感情が分けられている場合が多かったのでこの主成分が感情のグループ化に有効な情報を持っているのではないかと考え，第 2，第 4，第 5 主成分で 3 次元散布を作成し，感情別に点をつないだ (図 15) ．

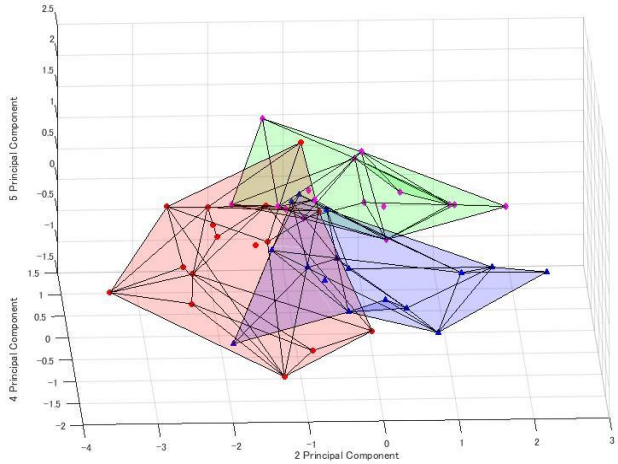

図 15.「怒り」「喜び」「悲しみ」3 次元散布図 (PC2 – PC4 – PC5)

図 15 を見ると，3 つの感情が大まかにグループ化できていた．上図の第 2 主成分 PC2 はピーク値の総合的な高さ，第 4 主成分 PC4 はピーク値の分散と解釈したが，第 5 主成分 PC5 は予測ができなかった．しかしこの結果から，フォルマントとピーク値からグラフを作成し,判別関数によって感情が認識できるのではないかと考えられる．

4. 各主成分と生理的な観点からの考察

図 16 に第 1 , 2, 4, 5 主成分における元の変数の寄与を示す．第 1 主成分が周波数の全般的なばらつきを表すのに対し,第 2 主成分は音量の全般的な大きさを表している．また，第 4 主成分は F1 周波数の低さ,と F2 の周波数の高さ，かつ高音成分の強度の強さを表している，第 5 主成分は低音成分の周波数の低さかつ，F1 周波数と高音周波数の強度の強さを表している。図 12 で見るように，第 1 主成分と第 2 主成分は寄与率が 70%以上と大きいが，感情の判別をするには不都合であることが分かった。

一方，感情の判別には第 2，第 4，第 5 主成分が有効であるようである．生理的な観点から感情を見ると，怒っている時は全体的に音量が大きい．従って、どのフォルマントのピーク値も高くなる．また，悲しい声を出す場合は下顎が下がり，それとともに舌，喉頭が下がる．一般に，発話中の舌の位置は第 1，第 2 フォルマントに反映され，第 1 フォルマントが舌の上下位置，第 2 フォルマントが前後位置に対応する[5]．よって，悲しい場合，第 1 フォルマントの周波数が低くなることが予想される．さらに，喜ぶ時には口蓋奥の口蓋帆が引き上がって高音域の周波数が高くなるとともに第 2 フォルマントの周波数が高くなる．

従って，怒りの音声の第 2 主成分が高い結果は妥当だと考えられる．また，悲しみについては第 1 フォルマントの強度が高い一方，高音域の強度が低く，第 4，第 5 主成分がそれぞれ負に現れるのと対応している．喜びでは第 1 フォルマントの強度が低い一方，高音域の音量が高くなっており，第 4, 第 5 主成分がそれぞれ正に現れるのと対応している．今後，音声感情を，今回のようなカテゴリ分けでなく，2 次元の感情平面に投影する際，第 4，第 5 主成分を用いることが有効と考えられる．

5. おわりに

本研究ではケプストラム分析で得た第 1 から第 6 フォルマントの周波数とピーク値を主成分分析で分析した結果，第 4、第 5 主成分が感情の判別に有効であることを得て，高音域のフォルマントが判別に影響していることが分かった．また，PC4 - PC5 平面での感情の配置と感情の円環モデルとの対応関係について期待ができる結果を得た．

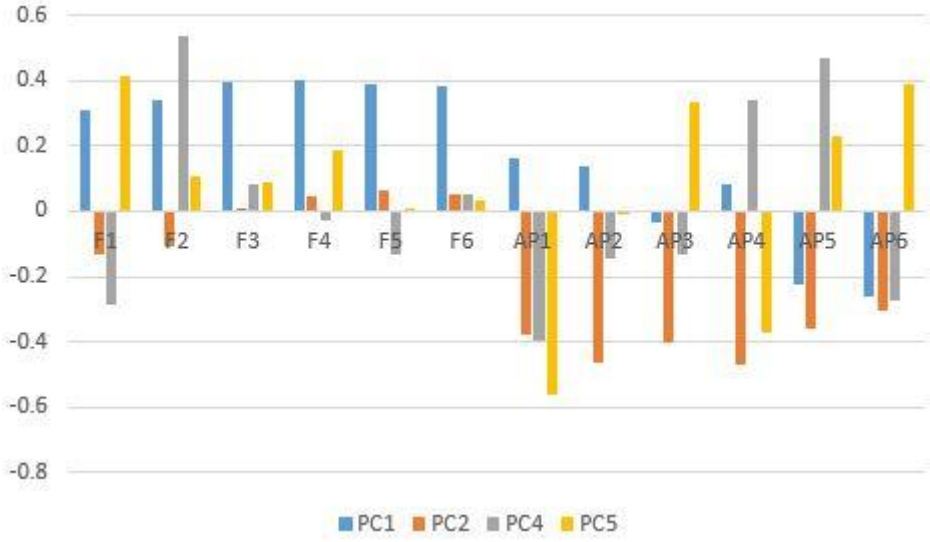

図 16. 各主成分とフォルマントの寄与

今後は，より詳細な判別分析を行い，感情音声の判別関数を作成すること， Arousal – Valence の 2 次元感情平面への対応づけを行い，声を人に聞かせて判定する官能検査と比較して，これらの数値的な手法の妥当性を評価することを計画している．

謝辞

本研究において，慶應義塾大学感情音声データベースを提供していただいた国立情報学研究所　音声資源コンソーシアムには深謝いたします．また本研究の一部は科学研究費補助金（基盤研究(c)課題番号 23611053）の助成を受けました．

参考文献

[1] 赤木正人: 音声に含まれる感情情報の認識 – 感情空間をどのように表現するか –, 日本音響学会誌，Vol.66，No.8，pp.393-398 (2010).
[2] 野田哲矢　矢野良和　道木慎二　音声を用いた話者適応型の感情認識に関する考察，第 22 回ファジィシステムシンポジウム公演論文集 6B4-3 (2006).
[3] J.A. Russell: A Circumplex Model of Affect, J. Personality and Social Psychology, Vol.39, No.6, pp. 1161-1178 (1980).
[4] 佐野智子: 不快感情の認知に影響を与える音響的指標，青山心理学研究，Vol.5，pp. 25-36 (2005).
[5] Tsubota, Y. et al.: Format Structure Estimation Using Vocal Tract Length Normalization for CALL Systems, Acoust. Sci. & Tech., Vol.24, No.2, pp.93-96 (2003).