怒りの感情音声における音響特徴量と感情知覚との関係に関する研究

平館 郁雄 (010095)

北陸先端科学技術大学院大学 情報科学研究科

2002 年 2 月 15 日

キーワード: 感情音声, 音響特徴量, アクセント部, Neutral, Cold Anger, Hot Anger.

1 はじめに

人間の最も基本的で日常的なコミュニケーション手段の一つとして音声がある。近年、音声を人間とコンピュータとのインターフェースとして用いることが実用化されつつある。この際、コンピュータと音声でコミュニケーションをとる時に、言語情報だけでなくパラ言語・非言語情報の伝達が行われれば、よりフレンドリなコミュニケーションが実現できる。その中でも特に、人間らしく聴きやすい合成音には感情情報が必要不可欠だと考えられる。

これまでの感情音声の研究においては、全発声区間という大局的な部分における平均的な変化を分析したものが多く、音響特徴量としては基本周波数・パワー・持続時間などが主に扱われ、それらについてはほぼ一致した見解が得られている。しかし、文章におけるアクセントの有無など局所的な部分を十分に分析したものは無い。

そこで、本研究ではアクセント部における音響特徴量と感情知覚との関係を調べることを目的とする。本研究は、「怒り」という感情を激しい怒りを表す「Hot Anger」と、押し殺した怒りを表す「Cold Anger」とに分け、「Neutral」の音声との比較検討を行った。音響特徴量としては、基本周波数・パワー・持続時間・フォルマント周波数・スペクトルを扱い、アクセントにおける変化に注目して分析を行った。

2 アクセントについて

韻律情報の中で、アクセントは合成音の高品質化の上で特に重要で、これには声の高さ・強さ・長さなどが関連するが、日本語では基本周波数がこれらを支配する直接的原因であるとされている [1]。日本語のアクセントは、高低の 2 レベルがあり、アクセントのあるモーラの直後にレベルが高から低に移る。これをアクセント核と呼ぶ。日本語の n

モーラ単語には、アクセント核が存在しない 0 型か、または核の位置が k=1(モーラ目) から k=n(モーラ目) までのいずれか 1 個所にある n 個のアクセント型が可能で、それぞれ k 型と呼ばれる。本研究ではこのアクセント型を基準として全文章に適合させた。

3　感情音声データ

3.1　感情音声データ

声優・演劇経験者は一般人に比べ、音声によって感情状態の表現を行う手法を的確に心得ている [2] ことがわかっているので、本研究で扱う感情音声データはプロの声優 (女性) から採取した計 179(9 パターン、20 文章) サンプルを用いた。9 パターンの感情は「平静」を表す「Neutral」が 1 種類と、「喜び」を表す「Joy」、「悲しみ」を表す「Sadness」、「押し殺した怒り」を表す「Cold Anger」、「激しい怒り」を表す「Hot Anger」が各 2 種類となっている。

3.2　聴取実験

本研究で扱う感情は、話し手側ではなく聞き手側に存在するものを指す。そして本研究では、感情情報を音声データの分析によって得るので、音声データが本研究で分析するのに十分有用であるかを判断するため、聴取実験を行った。

聴取実験は、採取した音声データを 1 文章ずつ 179 サンプル呈示し、その音声が「Neutral」「Joy」「Cold Anger」「Sadness」「Hot Anger」のどの感情のものなのかを判断させた。被験者は大学院生 9 名であり、その全てが正常聴力を有する。実験は、防音室内でヘッドフォンによる両耳受聴で行った。

聴取実験の結果、各感情において高い認識率のサンプルが多数存在しており、感情情報の抽出を目的とした分析にとって十分だと考えた。

4　音声データの分析

過去の研究ではアクセント部という局所的な部分の分析が十分行われていないため、本研究ではアクセント部に注目し音響特徴量の分析を行った。そこで、基本周波数・パワー・持続時間・フォルマント周波数・スペクトルといった音響特徴量を抽出・変換・合成するモデルを用いる必要があるので、この音声分析変換合成系として高品質な合成音声を作成できる STRAIGHT(Speech Transformation and RepresentationusingAdaptive Interpolation of weiGHTed spectrogram)[3] を採用した。本研究において、基本周波数・振幅・スペクトルといった音響特徴量は STRAIGHT を用いて抽出した。

4.1 基本周波数

基本周波数は、文章の有声区間における長時間平均とアクセントにおける変化率を分析した。アクセントにおける変化率は、アクセントレベルが低から高へ上昇する部分と高から低へ下降する部分のそれぞれについて、低レベルモーラの基本周波数に対する高レベルモーラの基本周波数の比を表す音韻変化率と、低レベルモーラから高レベルモーラへの時間に対する基本周波数の変動を表す時間変化率の 2 種類ずつを分析した。

有声区間における長時間平均を分析した結果、「Cold Anger」は「Neutral」よりも低い傾向があり、「Hot Anger」は「Neutral」よりも高い傾向を示した。

同様に、アクセントにおける変化率も分析した所、アクセントレベル上昇時は音韻変化率・時間変化率ともに、「Neutral」より「Cold Anger」は小さく、「Hot Anger」はかなり大きいという傾向が見られた。アクセントレベル下降時では、「Hot Anger」の時間変化率は「Neutral」よりも大きい傾向が見られたが、「Hot Anger」の音韻変化率と「Cold Anger」の音韻変化率・時間変化率は「Neutral」とほぼ同じくらいだった。

4.2 パワー

パワーに関しては STRAIGHT で振幅を抽出し、パワーの変化率を求めた。変化率は、アクセントレベル上昇時・下降時における音韻変化率・時間変化率について分析した。

その結果、アクセントレベル上昇時において「Cold Anger」は音韻変化率・時間変化率ともに「Neutral」とほぼ同じくらいであるのに対し、「Hot Anger」は「Neutral」よりもかなり大きな変化率を示した。アクセント下降時においては、「Cold Anger」「Hot Anger」は音韻変化率・時間変化率ともに「Neutral」と比較して大きい傾向が見られた。しかし、「Hot Anger」の変化率はアクセントレベル上昇時ほど大きくなかった。

4.3 持続時間

持続時間に関しては、文全体、アクセント、アクセントの母音・子音、非アクセント、非アクセントの母音・子音をそれぞれ測定し、「Neutral」に対する「Cold Anger」「Hot Anger」の比を分析した。

分析した結果、「Cold Anger」「Hot Anger」は文全体の持続時間において「Neutral」より短いが、文全体の中でも非アクセントよりアクセントで短くなっている傾向を示した。しかも、子音は各感情間でそれほど違いが表れず、特に母音で短くなっていた。文全体、アクセント、アクセントの母音、非アクセント、非アクセントの母音という持続時間において、「Cold Anger」は「Hot Anger」よりもさらに短い傾向が見られた。

4.4 フォルマント周波数

フォルマント周波数は、LPC により予測パラメータから推定した。分析は、アクセントレベルが低から高へ上昇する時の高レベルモーラにおけるフォルマント周波数を検討した。この LPC 法によるフォルマントトラッキングは SpeechTools [4] を用いて行った。

「Neutral」と「Cold Anger」のフォルマント周波数の関係をプロットしたものが図 1 である (縦軸、横軸はそれぞれ「Neutral」「Cold Anger」のフォルマント周波数 [kHz] を常用対数で表示したもの)。傾向を分かりやすくするため $y = x$ を波線で表示した。図 1 を見ると、全体的に「Cold Anger」のフォルマント周波数は「Neutral」のフォルマント周波数よりも低い傾向を示している。中でも、F1 において最も「Neutral」からのずれが見られ、次数が上がるにつれて「Neutral」との差は小さくなっていき、F4 ではほとんど差は見られなかった。

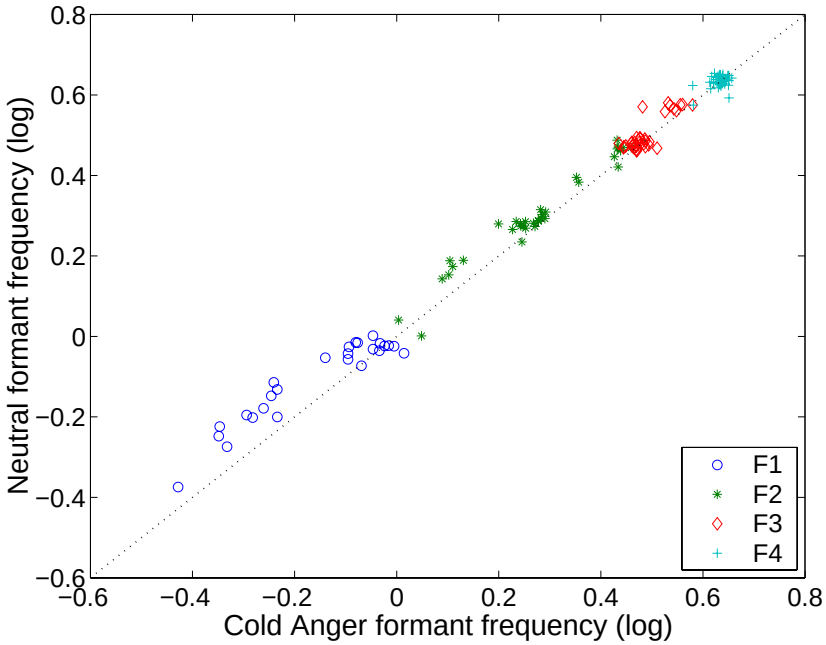

図 1: 「Neutral」と「Cold Anger」のフォルマント周波数の関係

同様に、「Neutral」と「Hot Anger」のフォルマント周波数の関係をプロットしたものが図 2 である。図 2 を見ると、全体的に「Hot Anger」のフォルマント周波数は「Neutral」のフォルマント周波数よりも高い傾向を示している。そして「Cold Anger」の時と同様に、F1 において最も「Neutral」からのずれが見られ、次数が上がるにつれて「Neutral」との差は小さくなっていき、F4 ではほとんど差は見られなかった。よって、本研究では F1〜F3 に注目し、プロットされた分析結果から最小自乗的に 2 次の近似式を導いた。そして、その近似式を合成音作成の際の周波数制御に適用することにする。

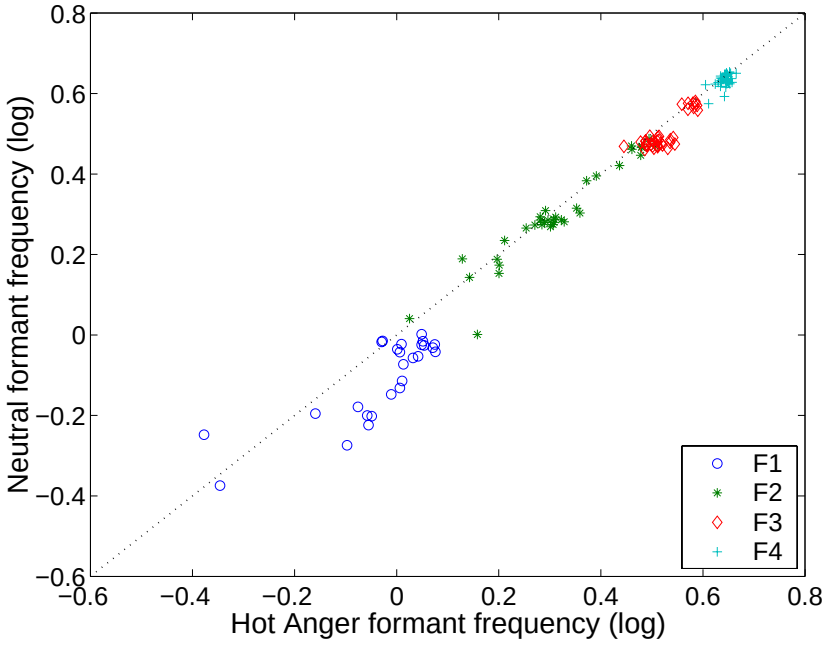

図 2: 「Neutral」と「Hot Anger」のフォルマント周波数の関係

4.5 スペクトル

スペクトル分析に関しては、アクセントレベルが低レベルから高レベルへ移るときの低・高レベルモーラにおけるパワースペクトルを分析した。そして、人間の内耳の基底膜の周波数特性を考慮した表現に対応させるため、周波数軸を ERB rate[5] に変換した。

「Neutral」「Cold Anger」「Hot Anger」という感情ごとのスペクトル概形の特徴を調べるために、35 ERB rate までを 5 ERB rate ずつ 7 つの帯域に分割し、スペクトル全体の平均パワーを基準としたときの各々の帯域におけるパワーを分析した。アクセント低レベルにおけるスペクトル分析の結果を表 1 に示す (カッコ内は同帯域の「Neutral」に対するパワー差)。

表 1: アクセント低レベルのスペクトル分析結果 (単位：[dB]

	0〜5 ERB rate	5〜10 ERB rate	10〜15 ERB rate	15〜20 ERB rate	20〜25 ERB rate	25〜30 ERB rate	30〜35 ERB rate
Neutral	13.6	25.4	20.1	10.3	2.8	0.1	-6.8
Cold Anger	17.3 (+3.7)	26.2 (+0.8)	19.8 (-0.3)	12.4 (+2.1)	3.5 (+0.7)	1.6 (+1.5)	-8.6 (-1.8)
Hot Anger	1.7 (-11.9)	15.1 (-10.3)	16.1 (-4.0)	11.6 (+1.3)	2.7 (-0.1)	4.2 (+4.1)	-7.5 (-0.7)

表 1 を見ると、「Cold Anger」は「Neutral」とほぼ同じ様なスペクトル概形を持っていると言えるが、「Hot Anger」は 「Neutral」と比較して 25〜30 ERB rate においてかなりパワーが強調されている。そして、「Cold Anger」「Hot Anger」ともに 30〜35 ERB rate においてパワーが減少している。

同様に、アクセント高レベルにおけるスペクトル分析の結果を表 2 に示す。表 2 を見ると、「Cold Anger」「Hot Anger」ともに 25〜30 ERB rate において「Neutral」と比較するとパワーが強調されている。アクセント低レベルの時と同様に、30〜35 ERB rate においてパワーが減少している傾向が見られた。

以上のアクセント低レベル・高レベルにおけるスペクトル分析の結果から、「Neutral」から「Cold Anger」「Hot Anger」の合成音を作成するためのスペクトル制御規則を導いた。

表 2: アクセント高レベルのスペクトル分析結果 (単位：[dB]

	0〜5 ERB rate	5〜10 ERB rate	10〜15 ERB rate	15〜20 ERB rate	20〜25 ERB rate	25〜30 ERB rate	30〜35 ERB rate
Neutral	9.5	22.8	19.9	11.5	4.7	-1.0	-6.6
Cold Anger	14.2 (+4.7)	24.7 (+1.9)	19.2 (-0.7)	12.4 (+0.9)	7.6 (+2.9)	2.2 (+3.2)	-9.9 (-3.3)
Hot Anger	-7.8 (-17.3)	3.4 (-19.4)	12.5 (-7.4)	10.8 (-0.7)	6.4 (+1.7)	3.7 (+4.7)	-6.8 (-0.2)

5 結論

「怒り」の感情音声における音響特徴量と感情知覚との関係を調べるために、基本周波数・パワー・持続時間・フォルマント周波数・スペクトルといった音響特徴量について、「Neutral」「Cold Anger」「Hot Anger」の感情間で比較した。

その結果、基本周波数に関しては有声区間の長時間平均とアクセントレベル上昇時に、パワーに関してはアクセントレベル上昇時に大きな特徴が見られた。持続時間に関しては、非アクセント部よりもアクセント部で、子音より母音で大きな特徴が見られた。フォルマント周波数は、F1 において感情間で顕著な差が見られ、次数が上がるにつれて差は見られなくなった。スペクトルに関しては、25〜30 ERB rate で「Cold Anger」「Hot Anger」は「Neutral」よりもパワーが強調されていることがわかった。

分析結果より、「Neutral」から「Cold Anger」「Hot Anger」の合成音を作成する規則を導いた。今後、この規則を用いて合成音を作成し、確認聴取実験を行う必要がある。

謝辞:
本研究において用いた感情音声データを提供・利用許可して頂いた (株) 富士通研究所の片江伸之氏に深謝する。

参考文献

[1] 古井 貞熙, “音声情報処理”, 森北出版株式会社,1998.

[2] 平賀 裕, 斉藤 善行, 森島 繁生, 原島 博, “音声に含まれる感情抽出の一検討”, 信学技法, HC93-66, pp.1-8, Jan.1994.

[3] 河原 英紀, “聴覚の情景分析と高品質音声分析変換合成法 STRAIGHT”, 日本音響学会講演論文集, 1-2-1,pp.189-192, Sep.1997.

[4] 天白 成一, 平原 達也 “UNIX 音声研究用ツール -SpeechTools-”, 日本音響学会講演論文集, pp.331-332, Mar.1991.

[5] B.R.Glasberg, B.C.J.Moore, “Derivation of auditory filter shapes from notched-noise data”, Hearing Research, 47, pp.103-138, 1990.