E-049

声優の発話の音響特徴量分析及び確率モデルの作成

Acouutic analysis of speech of radio actor and probabilistic model's making

原 雄太郎* 伊藤克亘†
Yutaro Hara Katsunobu Itou

1 まえがき

日本では、幅広い分野の様々なアニメーション作品が制作されてきた．この幅広さを支えてきた、一つの要素が、声優による声の演技である．この声の演技は「アニメ声」という言葉があることからもわかるよう非常に特徴的である．

本研究では，この特徴的なアニメーションの声優の音声表現のうち，特に感情表現を取り上げる．アニメーションの感情表現は，感情を表現するときに併用される身振りや手振りや表情が実写とは異なるため，実写とは異なっている．この点に着目し，本研究では，実写ドラマとアニメーションの両方が制作された作品を取り上げ，それらの発話を比較することで，アニメーション声優の発話の特徴を明らかにすることを目的とする．

これらの比較から，アニメーション声優の感情表現を区別するのに適した音響特徴量を明らかにする．さらに，それらの特徴量を確率モデルによりモデル化し，発話の感情の自動分類を行う．自動分類が可能になれば，アニメーション作品の発話コーパスの感情ラベルを自動付与することにより，感情表現コーパスの整備が期待できる．また，アニメーションのような感情表現の自動判定システムに応用すれば，感情表現の演技練習の支援も可能になり，アニメーション作品の制作の可能性を広げることも可能になる．

2 従来研究

アニメーション作品に用いられる発話は，特定の研究のためではなく，商業的利益のために使われている．このようなアニメーション作品に対する研究は極僅かではあるが，アニメ映画における声優の感情表現を研究しているものなどが存在する [1]．文献 [1] では，イントネーションは感情の表現に対して重要でないことや，発話の強度と感情には強く関連があることなどが明らかになっている．その他には，強い恐怖の感情表現について研究しているものが存在する [3]．

また感情分類に有効な特徴量に関する従来研究も多く存在する．文献 [6] では，言語に依存しない特徴量として，Teager Energy Operator(TEO) が用いられている．TEO は周波数のピーク部分を強調することができる．この特徴量は周波数ピークに敏感であるため，平静のような曖昧な感情の特徴を検出しにくい．また特徴量の時間変化は感情認識に有効であるといわれている．文献 [5] は，ドライバーの精神状態の情報を用いて安全性を提供するという目的で，自動車環境における感情認識を扱っている．この研究で分類される感情には，「怒り」，「嫌気」，「恐怖」，「喜び」，「平静」，「悲しみ」，「驚き」の 7 感情を扱っている．この文献では，スペクトル特性は強い音素と，発話の音声内容に依存すると述べられており，これは音響解析における内容の独立性の前提における欠点であるとも述べられている．また 200 以上の音響特徴の中から線形判別分析により，感情分類ための 33 次元の音響特徴セットがあげられている．

また感情認識に有効な分類器として [5] では SVM(Support Vecter Machine) をあげている．文献 [5] では kMeans, GMM(Gaussian Mixture Models), MLP(Multi Layer Perceptron), そして SVM によって分類を行っており，これら 4 つの手法の中で最も有効とされている手法が SVM であった．特に SVM の中でも最も有効な手法として ML-SVM(Multi Layer-Support Vector Machines) をあげており，これは 1 つのクラスが残るまで 2 クラス決定を繰り返すというものである

3 音響特徴量分析

本研究ではまずアニメーション声優の発話の音響的特徴を明らかにする．近年では，発話における感情を表現するような音響的特徴については，広範囲なパラ言語学情報に基づく特徴セットを用いている．[4] では，F0，強度，持続時間，声の質などが用いられており，[2] では言語学の情報が用いられている．

3.1 音声素材

本研究では，実写ドラマとアニメーションの両方が制作された作品の発話を比較することで，アニメーション声優の発話の特徴を明らかにする．そのため，上記の条件を満たしている日本アニメーションの「のだめカンタービレ」の DVD-Video から，発話音声の抽出を行った．音声サンプルはサンプリング周波数 48kHz，離散化ビット数 16 ビットで抽出した．前後の文脈を考慮し，発話内容が一致するような，109 の組み合わせの音声サンプルを用いた．例としては，あるキャラクターが同じ場面で，ドラマでは「テンポはゆっくりでいいから、このくらいで」という発話をしているが，アニメーションでは「テンポはゆっくりめでいいから、いくぞ」のように，発話内容は同じであるが発話自体は異なるようなサンプルがあげられる．また，この音声サンプルは感情情報に依存せずに収集した．

* 法政大学大学院情報科学研究科情報科学専攻,Hosei University
† 法政大学,Hosei University

3.2 分析

3.1 の音声素材を用いてアニメーション声優と実写俳優の発話を比較した．声の高さは F0 で分析した．発話のパワーは，異なる録音環境の影響を考慮して，約 21ms 毎のパワースペクトルの総和の最大値と，音声素材中のドラマとアニメの発話の最小パワーを求めて，それぞれの比で表す．単位は dB である．

3.2.1 発話の F0 に関する考察

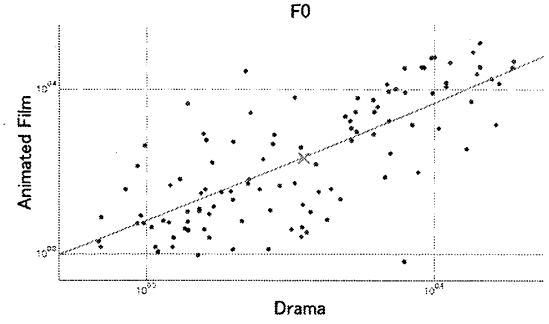

図 1. 声優と俳優の発話の F0 の比較

アニメーション声優の発話の平均 F0 は 213Hz であり，実写俳優の平均 F0 は 205Hz であった．また約 6 割のサンプルで，声優の F0 が俳優の F0 よりも大きい値であったが，図 1 の分布からは，声優と実写俳優の特徴量に大きな偏りはみられなかった．結果的に，声優と俳優の F0 からは明確な違いは得られなかった．しかし，相関係数が 0.75 という高い相関値を示しており，図 1 の分布からも，アニメーション声優の発話と実写俳優の発話の F0 には，強い相関関係があることがわかる．

このように高い相関関係がみられることから，同一のキャラクターには，実写でもアニメーションでも声の高さの傾向が一致するような話し方をしていると考えられる．また同じ場面や発話内容でも，声優と俳優とで異なった感情で演じている発話が存在した．部屋が汚いことを注意する場面で「掃除」と発話しているが，ドラマでは「怒り」の感情で発話しているのに対して，アニメーションでは「悲しみ」の感情で発話しているような例があげられる．このような感情の違いによって，その発話の特徴量が影響を受けた可能性がある．

3.2.2 発話のパワーに関する考察

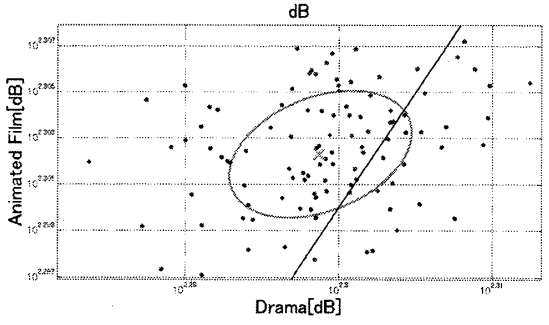

図 2. 声優と俳優の発話のパワーの比較

アニメーション声優の発話のパワーは平均 101.4dB，実写俳優の発話のパワーは 97.8dB であった．図 2n の分布から，声優の方が発話のパワーが大きい傾向にあると考えられる．

しかし，同じ録音環境で収録した発話であっても，収録後に行う音声処理によってパワーが補正されてしまっている可能性がある．そのため，特徴量として得られたパワーが，人間ではなく録音環境に依存している可能性が考えられる．以上のことから，声優の発話のパワーが大きいとは一概には言えないかもしれないが，約 8 割のサンプルにおいて声優の発話のパワーが俳優のそれを上回っていることから今回の結果にはある程度の信頼性があると考えられる．

4 感情の確率モデル

4.1 確率モデル

3 章で評価した特徴量などを，確率モデルによりモデル化し，発話の感情の自動分類を行う．各音響特徴量 (F0，ダイナミックレンジ，ΔF0，デシベル) の分布を正規分布と仮定し，それらの平均と共分散を求めることで感情の自動判別のための確率モデルを作成した．F0 は F0 値を話者ごとに正規化した値，ダイナミックレンジは最大 F0 を最小 F0 で割った値，ΔF0 はフレーム毎の F0 の差分の絶対値の平均値，デシベルは，約 21ms 毎のパワースペクトルの総和の最大値と，全サンプル中の発話の最小パワーの比で表わす．

4.2 評価実験

F0，ダイナミックレンジ，ΔF0，デシベル，以上の 4 つの特徴量で音声サンプルを評価し，感情ごとの音響特徴量の違いを明確にする．

本研究では、一つに分類される感情を，細かい感情に分類した．これは感情の度合いによって，特徴量が大きく異なる可能性があるからである．怒りの感情を，「内に込めた怒り」と「爆発させる怒り」の 2 つに区別し，後者の「爆発させる怒り」を怒りとして定義することにする．悲しみも「嘆き叫ぶ悲しみ」と「内に込めた悲しみ」の 2 つに区別し，後者を悲しみとして定義した．

学習データには，「喜び」，「怒り」，「悲しみ」，「平静」の 4 感情をそれぞれ 25 個ずつ，のべ 100 個のサンプルを特徴量を用いた．また，学習データに用いられていない音声サンプルを，声優 5 人について感情ごとに約 10 個ずつ，のべ 180 個程度のサンプルの特徴量を評価データとして用いた．

確率モデルには感情が分類されていない感情発声を入力し，各感情に対する尤度を求める．

$$Motion = \arg\max_{i}[p(x \mid C_i)] \tag{1}$$

式 (1) によって尤度を求め，尤度が最も大きかった感情に分類する．

今回の実験では F0 情報のみを含む 3 次元確率モデル (F0，ダイナミックレンジ，ΔF0) と，パワー情報も含む 4 次元確率モデル (F0，ダイナミックレンジ，ΔF0，デシベル) の 2 つのモデルに関しての感情分類結果を考察する．また各モデルの結果を比較することで，パワー情報の有効性についても考察する．

4.3 音声素材

3.1 章と同様に確率モデルの作成には，日本アニメーションの「のだめカンタービレ」の音声素材を用いている．確率モデルに用いる音声サンプルには，4.1 で述べた声優 5 人分の感情発声サンプルを約 280 個を用いている (学習データは約 100 サンプル，評価データは約 180 サンプル)．収集した感情発声サンプルは主観のみで分類するのではなく，客観性を持たせるため 2〜3 人で分類をしている．

4.4 3 次元モデル

4.4.1 結果と考察

確率モデルにより，評価データに対する感情ごとの認識結果が得られた．表 1 に 3 次元での各感情音声の認識結果を示す．なお，横が入力した感情，縦が出力された感情である．

	喜び	怒り	悲しみ	平静
喜び	53%	30%	0%	17%
怒り	36%	58%	0%	9%
悲しみ	10%	0%	88%	2%
平静	16%	0%	24%	60%

表 1. 3 次元モデルの認識結果

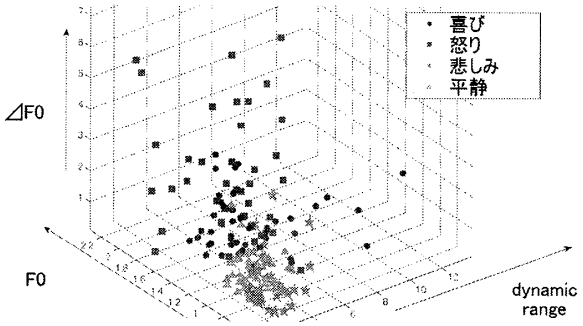

図 3. 音響特徴量の 3 次元グラフ (●=喜び，■=怒り，★=悲しみ，▲=平静)

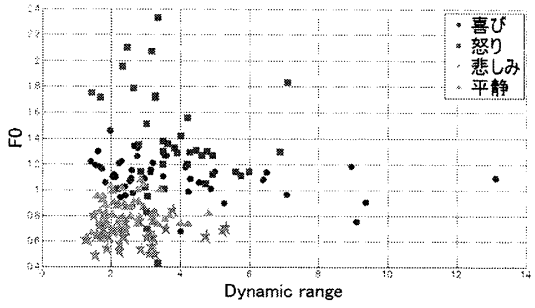

図 4. F0 とダイナミックレンジの 2 次元分布 (●=喜び，■=怒り，★=悲しみ，▲=平静)

表 1 によると，悲しみの認識率が 90% 近くもあり，最も認識率が高かった．怒りと平静は悲しみに次いで約 60% の認識率となった．また喜びは最も低い認識率となったが，認識率は 53% であり半数以上の喜びの評価データを正しく認識することができた．また表 1 より，喜びは怒りと平静に誤認識されやすく，悲しみには誤認識されにくかった．怒りは喜びに多く誤認識されやすかったが，悲しみには誤認識されにくかった．悲しみは最も高い認識率であり，他の感情に誤認識されにくく，特に怒りには誤認識されにくかった．平静はやや悲しみに誤認識されやすい結果となったが，怒りには誤認識されにくかった．以上の結果から，喜びと怒りは悲しみに分類されにくく，悲しみと平静は怒りに分類されにくいと考えられる．

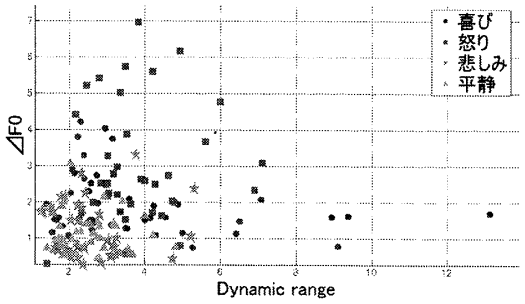

図 5. F0 とダイナミックレンジの 2 次元分布 (●=喜び，■=怒り，★=悲しみ，▲=平静)

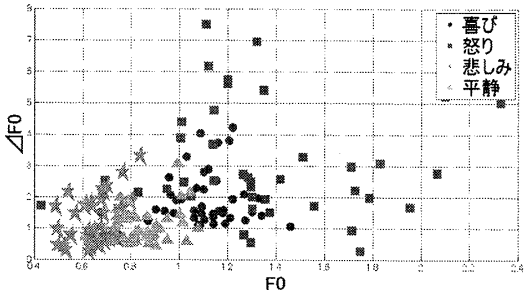

図 6. F0 と F0 の 2 次元分布 (●=喜び，■=怒り，★=悲しみ，▲=平静)

また分類された感情発声の音響特徴量の分布を，2 次元及び 3 次元グラフで比較する．図 3 のように，分類されたサンプルの音響特徴量を 3 次元グラフにすると，悲しみと平静の特徴量の分布の重なりが顕著であった．これが表 1 において，平静が悲しみに誤認識されやすかった原因の一つであると考えられる．また，怒りと喜びは大きい値の分布をとっていたが，悲しみと平静は小さい値で分布されていた．

図 4 は，ダイナミックレンジと F0 の 2 次元分布図である．図 5 のダイナミックレンジと ΔF0 の分布図と比較すると，図 4 や図 6 の組み合わせの分布では，感情ごとの分布の重なりが小さいことがわかる．そのため F0 は，ΔF0 やダイナミックレンジと比較して，感情認識に用いる特徴量としては強い特徴量であると考えられる．

表 2 から，喜びと怒りのダイナミックレンジがほぼ同等の値をとっていることがわかる．これが喜びと怒りが相互に誤認識されやすかった大きな要因と考えられる．また表 2 から，誤認識されにくい感情間では，音響特徴量の値に大きなひらきが存在することもわかる．特に喜びと怒りの各特徴量が，比較的大

	喜び	怒り	悲しみ	平静
Dynamic range	3.79	3.8	2.5	2.58
F0	1.1	1.36	0.65	0.84
ΔF0	1.9	3.05	1.24	1.07

表 2. 感情発声の音響特徴量の平均値

きい値を示しているのに対し，悲しみと平静の各特徴量は比較的小さい．表 1 から喜びと怒りは悲しみに分類されにくく，悲しみと平静は怒りに分類されにくいという結果を得たが，上で述べたことが大きな要因であると考えられる．

4.5 4 次元モデル

4.5.1 結果と考察

表 3 にパワー情報も含めた 4 次元での各感情音声の認識結果を示す．

	喜び	怒り	悲しみ	平静
喜び	25%	41%	18%	16%
怒り	19%	75%	0%	6%
悲しみ	7%	7%	76%	10%
平静	12%	16%	33%	39%

表 3. 4 次元モデルの認識結果

4 次元モデルは，3 次元モデルと比較すると，全体的に認識率が悪くなっていることがわかる．喜びの認識率が 25% と最も低くなっており，3 次元モデルと比較してもかなり認識率が低下している．また悲しみと平静についても認識率は低下しており，3 次元モデルと比較して，どちらも 20% ほどの認識率の低下がみられた．怒りは 75% と 3 次元モデルに比べ，認識率の大きな向上がみられた．しかし，表 3 から喜びが怒りに誤認識される確率は 41% と非常に大きい値となっており，喜びの特徴量の分布が怒りの特徴量の分布に吸収されてしまっていると考えられる．

結果的に，3 次元モデルと比較して悪いモデルになってしまっており，これは特徴量にパワー情報を用いたことが原因であると考えられる．アニメーションでは場面ごとに，キャラクター同士の距離などを考慮して，パワーを増幅させたり減衰させるなどの処理をかけている可能性がある．このような処理をかけると，一定の条件で録音した音源であっても，パワー情報から感情分類のための安定した特徴量が得られないと考えられる．そのため，アニメーションから切り出した発話サンプルから，パワー情報を用いて確率モデルを作成することは不適当ではないかと考えられる．

5 まとめ

アニメーション声優と実写俳優の発話の F0 を比較した結果，特徴量に大きな偏りはみられなかったが，強い相関関係がみられたことから，声の高い俳優に対しては，同じく声の高い声優を起用していると考えられる．アニメーション声優と実写俳優の発話のパワーを比較した結果，分布はアニメーション声優の方に偏っており，声優の方が俳優の発話の F0 よりも大きい傾向が得られた．

3 次元確率モデルでは悲しみの認識率が 90% と最も高く，最も低い喜びの認識率も 50% を超える結果となり比較的良好な結果であった．特徴量の分布では，悲しみと平静の分布のばらつきが小さく，喜びと怒りの分布のばらつきが大きいという結果が得られた．またパワー情報も含めた 4 次元モデルでは，3 次元モデルに比べ認識率が低下した．アニメーションから切り出したサンプルで，確率モデルにパワー情報を用いることは不適当である可能性がある．パワー情報を確率モデルに用いる場合には，実際に声優の声を録音するなどして同環境上での音声サンプルを用いなければならない．

今回はサンプルとしてアニメーションから切り出した音声を使用したが，今後は同環境上で録音された声優の感情発声を用いることで，パワー情報の有効性も検討していきたい．また感情分類に有効な特徴量でモデルを作成するとともに，より感情認識に適した分類器を用いて研究を進めていく．今回は 1 つの作品のみで分析を行ったため，データの増加によって得られた結果が大きく変動するかもしれない．今後はジャンルの違うアニメーションや発話サンプルを増やす必要があると考える．

参考文献

[1] N. Amir and R. Cohen. Characterizing emotion in the soundtrack of an animated film: Credible or incredible?, 2007.

[2] A. Batliner, S. Steidl, B. Schuller, D. Seppi, and more. Combining efforts for improving automatic classification of emotional user states. *Proceedings of IS-LTC 2006,Ljubliana*, pages 240–245, 2006.

[3] C. Clavel, I. Vasilescu, L. Devillers, T. Ehrette, and C. Richard. Fear-type emotions of the safe corpus:annotation issues. *Proc. 5th Int. Conf. on Language Resources and Evaluation (LREC)*, 2006.

[4] M. Schroder, R. Cowie, E. Douglas-Cowie, M. Westerdijk, and S. Gielen. Acoustic correlates of emotion dimentions in view of speech synthesis. *Proceedings Eurospeech*, pages 87–90, 2001.

[5] B. Schuller, G. Rigoll, and M. Lang. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine - belief network architecture. *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, 1:I–577–I–580, 2004.

[6] 直井克也, 松本哲也, 竹内義則, 工藤博章, and 大西昇. 感情に関する特徴量の検討. 信学技報, pages 37–42, 2005.