キャラクター音声のステレオタイプ識別のための音響分析

石井　沙季
Saki Ishii
法政大学大学院情報科学部ディジタルメディア学科
saki.ishii.5h@stu.hosei.ac.jp

Abstract

In Japan, subculture contents become popular. Also, a wide variety of characters were born. Many characters have typical character and behavior. Rocoguizing these has the advantage of making it easier to structure and understand the story. Therefore, the stereotype property of the character sound is clarified on the acoustic feature surface. The paralanguage information accompanying the voice seems to show the characteristics of the change in tone color, speech speed and pitch. From the results of the analysis of feature quantity, perform objective evaluation in the system based on maximum likelihood method. The batakaria distance using the proposed feature amount was the most distant, which was 31. In the evaluation by the maximum likelihood method, things using three speech speeds, F 0, LPC had the highest overall accuracy, 0.4463. Therefore, it is possible to distinguish stereotypes of male characters in the proposed feature quantity.

1　はじめに

現代の日本では、アニメやゲーム、舞台といったサブカルチャーコンテンツが人気である。アニメは年間 200 本ほどの新作がテレビやインターネットの動画サイトで配信される。人気の作品になると様々なメディアで展開され、声優の他にも役者がキャラクターを演じるため一人のキャラクターに対していくつもの声がついたりもする。このように、多種多様なキャラクターが数多く生まれてきている。

アニメなどでは、キャラクターの姿を見なくてもしゃべっている音声を聞くだけでそのキャラクターの言動や役割を推定できる。そこでキャラクター音声のステレオタイプ性について注目する。言語学の分野では言語内容からキャラクターのステレオタイプを明らかにする研究が行われてきた。実際にお嬢様や博士などの言葉遣いの特徴が指摘されている。言葉遣いや話し方、つまり音声でどのようなキャラクターであるかわかると物語に入りやすくなるという利点がある。本研究では、言語内容ではなく音声に着目し、キャラクター音声のステレオタイプ性を音響特徴面で明らかにすることを目標とする。音響的な特徴が明らかになると、音声合成の分野で役に立つのではないかと考える。

2　キャラクターのステレオタイプ性

ステレオタイプ [1] とは社会心理学、社会言語学の概念である。性別や年齢、容姿などの特徴で人間を分類し、そこに属するものが共通して持つとされるものがステレオタイプである。

多くの人は幼少期から物語の中でお姫様や王子様、ヒーロー、悪役などの様々なキャラクターを目にしている。上記のキャラクターに関しては立場や言動を無意識に決めつけているところがあると感じる。このように無意識に役割等を断定しているキャラクターがステレオタイプなキャラクターであると考える。

また、言語学の分野で役割語というものが研究されている。ある特定の言葉遣い（語彙・語法・言い回し・イントネーション等）を聞くと特定の人物像（年齢、性別、職業、階層、時代、容姿・風貌、性格等）を思い浮かべることができるとき、あるいはある特定の人物像を提示されると、その人物がいかにも使用しそうな言葉遣いを思い浮かべることができるとき、その言葉遣いを「役割語」と呼ぶ [2]。

これらを用いて言語内容からキャラクターのステレオタイプ性を明確にする研究が行われている。上記の内容を参考にし、キャラクター音声のステレオタイプ性を判断する。

以下に役割語の例を示す。太字になっているところが役割語である。ポケットモンスターの登場人物オーキド博士は「…、野生のポケモンを戦わせ、勝つと採集できるんじゃ…」[3] と話す。これは、博士が一般的に話すとされる＜博士語＞である。

また、エースをねらえ！の登場人物のお蝶夫人は「あたくしも賛成**ですわ**/音羽さんはいつも選手で実力があって」[4] と話す。これは、典型的な「お嬢さまことば」である。

このように役割語を用いてそれぞれのキャラクターらしい言葉遣いによってキャラクターのステレオタイプ性を明確にする研究が行われている。

3　パラ言語情報とキャラクター

キャラクターのステレオタイプの明確化は先に述べたように言語学の分野で行われている。ここでは役割語に引きずられてステレオタイプ性を決定しないようにするため、音声によって分類を行っていく。例えば音声には性別、年齢、態度、性格、感情、調子などが現れる。これは言語内容とは別の情報であり、パラ言語情報と呼ばれるものである。キャラクターには言葉遣いだけではなく、話し方にも個性がある。語尾の上昇、下降、弱弱しい声、とても力強い声など様々なものがある。以下に三つのセリフ [5] を載せる。

- 「環さまがあなたを構うのは、育ちが珍しいからよ」
  - 若い女性、気品がある、性格きつそう、皮肉こめてる、重い
- 「まあ、それが噂の？」
  - 若い女性、気品がある、感嘆、疑問、軽い
- 「貧乏な方は暇がないので豆も挽けないというのは本当でしたのね」
  - 若い女性、気品がある、驚き、若干の興奮、軽い

ここで挙げた文章はいずれもお嬢様のセリフである。セリフの下に各音声から得たパラ言語情報を記している。ここからお嬢様である音声は上のようなパラ言語情報を所持していると考えることができる。共通して得た情報は、「若い女性」「気品がある」であった。他に共通しない情報があることから音がついていると表現の方法が広がるということが推定できる。つまり、パラ言語情報の一部がキャラクターに関係していると考えられる。これらの情報は音声から感じる言語情報以外のものであると記した。したがって、音響的に特徴があると思われる。しかし、パラ言語情報と音響的な特徴の関連が明確になっているものは少ない。関連性がありそうな特徴量について見ていく。

Supervisor: Prof. Katunobu Itou

4　音響特徴量

従来のパラ言語と音響の関係の研究 [7] では基本周波数やパワー、持続時間などの韻律的特徴を利用したものが多い。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 [?]。これらから本研究では音色、F0、話速について分析していきたい。

4.1　音色

音声に抱く印象が音色に関係してくると思われる。スペクトル包絡の形に違いが現れることがわかっている。包絡とはスペクトルの大まかな形を取ったものであり、声道特性を表している。一人ひとり形が異なるもので、声色がこれによって定まる。どの母音においても包絡の形は言語情報を含む低域以外は同様の傾向を示すため、全母音の平均化スペクトル包絡で比較を行う。また全ての母音に対応させるためでもある。本研究では 21 次元の線形予測係数で求める。最小二乗的に予測誤差を最小にすることで係数が推定される。どの周波数帯域で違いが出るか、ホルマントの位置はどうか、強さはどうかなど特に特徴が出ている面を見ていく。第一、第二ホルマントは言語情報を含んでいるため男女では周波数の位置に違いは出るが、同一の性別ではあまり変わらないことがわかっている。音声的な特徴が出るのは第三ホルマント以降になる。

ここですべて同一の声優が演じた複数の年代・性別のスペクトル包絡を図 1 に示す。音声のバリエーションは「通常」「おばあさん」「男」「少年」「少女」「幼児」の 6 種類である。全て同一人物が発した音声とは思えないほど包絡の形に違いが出ている。同一人物の音声であってもはっきりと特徴が現れることが見て取れた。これは呼吸方法がかなり異なっているからであると考えられる。したがって、同一の声優の音声であってもキャラクターのステレオタイプが異なれば音色に特徴が現れるのである。

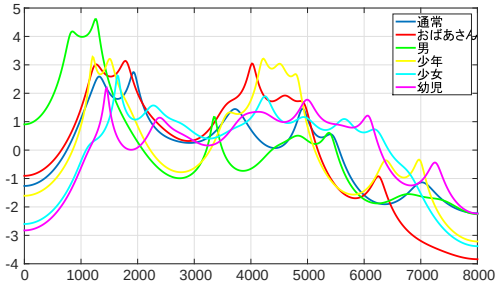

図 1　同一声優の「あ」のスペクトル包絡

似ている部分は分析に使う必要性がないため、特定の周波数帯域に注目していきたいと考える。LPC のパワーについても見ていく。

4.2　基本周波数 (F0)

F0 を推定することで基本周波数（音高）がわかる。音声スペクトルの自己相関を取ることで求めることができる。自己相関の式は以下の通りである。

$$r(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_{-T}^{T} x(t)x(t+\tau)dt \qquad (1)$$

男性キャラクターと女性キャラクターは性別が異なるため、違いが出ることは性別判定の研究からも明らかである。年齢が異なっても高さに違いが出ることがわかっている。同性のキャラクターであってもステレオタイプごとに特徴が現れるのではないかと考える。

また、音声の抑揚についても見ていく。音声の抑揚は F0 の前後の変化の差分を取り、その差分が大きい箇所をピークが現れているとする。ピーク数を音声長で割ることで抑揚を定義する。ピーク数が多いほど F0 の変化が多いということになる。

4.3　話速

お嬢様は全体的にゆっくりめ、真面目はきびきびとした一定の速さ、不良は激しく速いといった印象があることが考えられる。しかし、同様のステレオタイプであってもキャラクターごとに性格があるため、話速は異なる可能性もある。また、平均的な話速では明確な違いが出ない場合も考えられる。語尾の速度など、限定的な部分でも見ていく。話速の求め方で有名なものは二つある。音韻ごと、パワー変動である。音韻ごとはよく話速を求めるときに使用される方法で、発話内容に左右される [6]。一方、パワー変動を用いた方法では発話内容によらない話速を求めることが可能である。一定の時間長の中にどれだけのピークが現れるかで話速を定義する。今回はパワー変動を用いた方法で話速を求める。

図 2 はお嬢様なキャラクターのスペクトルと音声波形、ピークを重ねて表示したものである。単位時間あたりのピークの数から話速を求める。台詞は「環さまがあなたを構うのは、育ちが珍しいからよ [5]」である。図 2 での平均の話速は 3.09、冒頭の話速は 1.19、終盤の話速は 5.94 であった。図 2 を見てみると冒頭はピークの数が少なく、終盤は多くなっている。ピークの数が多ければ、それだけ音韻が現れていると考えることができるため話速が早くなる。逆にピークの数が少なければ、音韻が現れていないということであるため話速は遅くなると考えられる。

Δ MFCC のノルムを計算し、そのピーク数を時間長で割ることで求めた。平均の話速は単純にフレーズ全体で現れるピーク数を全体の時間長で割ればよい。しかし冒頭と終盤の話速はどこまでの長さにするか決めなければならない。そこで、本研究では冒頭の二割、終盤の二割をそれぞれの位置の話速と定めた。

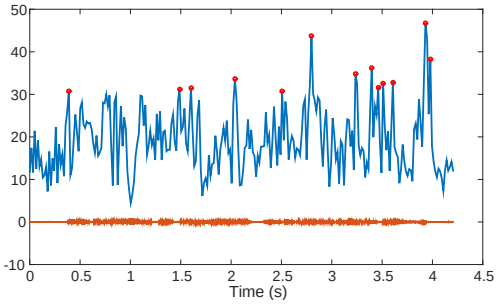

図 2　お嬢様の話速

5　聴取実験によるステレオタイプ性を持つ音声の選定

聴取実験により 3 つのステレオタイプ性を持つ音声を定義する。明らかにする対象のステレオタイプは「お嬢様」「真面目」「不良」である。対象のステレオタイプの選定基準は比較的音声が集めやすいと感じたものである。

5.1　音声データ収集

アニメ、ドラマ CD から音声を集めていく。収集する際、極端な感情音声（誰が聞いても怒っていると感じるものなど）は除く。これは感情を考慮すると細分化されすぎてしまう可能性があるからである。

基本的にそのキャラクターの通常時の声音を集めていく。BGM がついているとそちらに気を取られてしまい、正確な判断や分析を行えない可能性が高い。そのため BGM がついていない音声を集める。

5.2　音声の逆再生

収集した音声には言語情報が含まれている。本研究では音声のステレオタイプ性を明らかにすることが目的であるため、言語情報が含まれていると役割語 [2] などにより、音声を聴いた者が正確な判断を行うことが難しくなると考えられる。そこで言語情報を無くすために逆再生を施す。

逆再生の処理を行うと日本語をローマ字表記したものを逆から読んでいるような音声になる。そのため、何と言っているのか判断しにくくなるのである。逆再生をしたところで音響的な特徴には大きな違いは現れない。含まれている子音や母音は同一のものであるからである。しかし、聴こえ方の速さに違いが生じることが確認されている。話速に関しては、話し始めと話し終わりの速さが聴いた際の印象に影響する。そのため逆にすると異なった印象を抱く可能性があるのである。逆再生すると失われる情報があることは聴取時に感じる速度が変わることから明らかである。その点も踏まえ、逆再生した音声と通常再生する音声の二種類を用意する必要があると考えられる。

5.3　聴取実験

実験に使用した音声は通常再生音声と逆再生音声の二種類である。男性キャラクター音声のみを集めた。音声は無作為にアニメやドラマ CD からキャラクターの通常発話を収集した。音声には同一声優の別キャラクター音声を含んでいる。なぜアニメやドラマ CD から収集したかと言うと、音声によるキャラクターがはっきりしているからである。ドラマなどでは俳優がセリフだけではなく身体を使った演技によっても自身が演じるキャラクターを表現している。そのため、あまり音声にはっきりした特徴が現れないと考えられる。しかし、声優は声のみでどのようなキャラクターであるかを伝えなければならないため、そのような訓練を受けている。よって声優の演じた音声であるアニメ、ドラマ CD が適していると考えた。

聴取用の音声は動画として用意し、被験者には耳を覆う形のヘッドフォンで聞いてもらった。次の 16 種類から各音声にラベル付けをしてもらった。不良、優等生、熱血、中二病、オタク、少年、ヘタレ、天然、お兄ちゃん、俺様、純情、不器用、天真爛漫、中性的、ヒーロー、チャラ男。属性の選択にはアニメ視聴者の女性を対象としてアンケート調査を行った、女性が好きな男性キャラクターの属性人気ランキングの紹介記事を参考にした。

実験条件は表 1 に示した通りになる。

表 1　実験条件

男性キャラクターボイス	52（通常再生、逆再生）× 2=104 音声
キャラクター属性選択肢	16 種
被験者：大学生男女	6 名（男性 3、女性 3）

選択された音声の上位 3 つをステレオタイプ性を持つ音声だと決定した。また音声は 3 人以上が選択したものを収集した。集計の結果、「俺様」「優等生」「少年」を分析の対象とした。

しかし、どの音声も言語内容に左右されている可能性が出てきた。また、音声の選択にはかなりのばらつきがあった。被験者からの感想として、当てはまる選択肢がないように感じる音声もあったと言われた。選択肢に偏りがあるように感じるとも言われた。選択肢の改善が必要であると思われる。

6　音響特徴：俺様、優等生、少年

俺様、優等生、少年において以下のことが明らかになった。

6.1　音色

決定された各ステレオタイプの音声において母音部分を切り出し、平均化した LPC において比較を行った。図 3 を見ると 6100Hz から 7100Hz の帯域において包絡の形に違いが現れた。その帯域の振幅の強さは上から順に優等生、俺様、少年である。線形予測係数は声道を音響管の連接ととらえ、その特性を推定しているものである。少年は声帯を細かく弱めに振動させることで高く、幼めな印象を抱かせるのだと考えた。

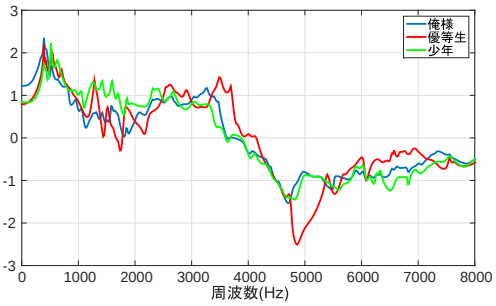

図 3　俺様、優等生、少年の 5 母音平均化スペクトル包絡

6.2　F0

結果を表 2 に示す。

表 2　F0 結果

	俺様	優等生	少年
F0(Hz)	133	125	184
抑揚 (個/秒)	3.20	0.08	1.85

F0 は低い順から優等生、俺様、少年になった。年齢層が低い方が基本周波数は高いということが知られている。これは人間の声の高さには声帯の長さが関係しているからである。人の声の高さは声帯の振動数によって決まる。振動数が多いほど高い声ということになる。振動数は振動するものの長さ、質量、緊張度などによって決まる。長いもの、重いもの、緊張が緩いものの方が振動数は少なく、音は低くなる。人は誰でも高い声から低い声まで出すことができる。これは声帯の厚み（振動部分の質量）や緊張度を筋肉の動きによって変えることで調節している。したがって、少年声を出すときは緊張度をきつくし、声帯の振動数を上げていると考えられる。また、俺様は音声によって極端に F0 が高くなるところが見られたため、フレーズ全体の F0 の平均を取った際に優等生よりも高くなったものと考えられる。

抑揚に関しては、少ない方から順に優等生、少年、俺様という順番になった。優等生ははっきりとした声で淡々と話すことで相手に対して言葉を聞き取りやすいように話しているのではないかと考えた。俺様は抑揚をつけることで相手に自身の要求を飲ませようとするのではないかと考えた。

6.3　話速

結果を以下の表 3 に示す。

表 3　話速結果

	俺様	優等生	少年
平均 (個/秒)	6.99	5.75	4.31
冒頭 (個/秒)	7.81	2.62	6.70
終盤 (個/秒)	3.90	10.46	4.04

フレーズ全体の話速の平均は遅い方から順に少年、優等生、俺様という順になった。俺様が速いのは音を聞いていると明らかである。

冒頭の話速は遅い方から順に優等生、少年、俺様であった。逆に終盤の話速は遅い方から順に俺様、少年、優等生であった。

発話全体の話速の印象は冒頭と終盤が関係していると考えられる。発話の冒頭が速く、終盤が遅いということは先に勢いをつけて相手を威嚇し、徐々にどすの効いた声にすることで自信の主張を通そうとしているのである。これが俺様である。逆に発話の冒頭が遅く、終盤が速いことは先に自身が何を言っているか相手に把握しやすくさせるためだと考えられる。ゆっくり話始めれば相手は何を話していたか聞き取りやすい。そして終盤は伝えることは伝えたという状態になるため早口になるのであろうと考えた。これが優等生である。

7 評価

7.1 バタチャリア距離による有効性の評価

ステレオタイプごとに学習データを作成し、特徴量の有効性についてはバタチャリア距離を用いて調べる。バタチャリア距離とはどれだけ分布間の距離が離れているかを調べるもので、距離が遠いものほど異なっていることを表している。したがってそれぞれの特徴量においてステレオタイプごとの距離が遠いものほど有効なものであると考えることができる。定義の式は以下の通りである。

$$\begin{aligned} BD(P_a, P_b) &= -\log_e \int_{-\infty}^{\infty} \sqrt{P_a(x)P_b(x)}dx \\ &= \frac{1}{8} u_{ab} \left\{ \frac{\sum_a + \sum_b}{2} \right\}^{-1} u_{ab}^t \\ &\quad + \frac{1}{2} \log_e \left(\frac{|\sum_a + \sum_b /2|}{|\sum_a|^{\frac{1}{2}} |\sum_b|^{\frac{1}{2}}} \right) \end{aligned} \tag{2}$$

使用した特徴量は A. 話速平均、B. 話速冒頭、C. 話速終盤、D.F0、E. 抑揚、F.LPC、G.LPC パワーの七種類（以下それぞれ A〜G）である。なお、モデル作成に使用した混合ガウスモデルの混合数は 2 である。結果を以下の表 4 に示す。

表 4　ステレオタイプ間のバタチャリア距離

	俺様, 優等生	俺様, 少年	優等生, 少年
A〜G(7)	27.5684	27.3640	31.4510
A〜F(6)	17.2359	25.3923	17.1136
A〜D,F,G(6)	12.0696	13.2536	11.6830
A〜C,E〜G(6)	10.4503	12.5207	27.2116
A〜D,F(5)	10.0715	25.5038	29.2079
A,D〜G(5)	14.0905	25.7551	28.6384

最も遠いものは今回提案した特徴量をすべて使用したときの優等生と少年であった。F0 の平均はおよそ 60Hz ほど離れており、スペクトル包絡の形に関しても優等生と少年は最も遠く、話速に関しても 3 つの項目すべてにおいて 1.2 から 1.7 倍ほどの差があった。そのため距離が遠くなったと考える。よってすべての特徴量が有効そうであると考えられる。

7.2 最尤法による評価

最尤法によって音声データの評価を行った。結果は表 5 の通りである。

全体としてはおよそ 4 割の精度を得ることができた。優等生においては他の二つに比べて著しく精度が低い。全くでない組み合わせもあった。これは最尤法にて評価する場合に特徴が俺様や少年にかぶる部分が多いためであると考える。しかし比較的特徴がはっきりしていた話速の三種で見てみると優等生の精度が最も高くなった。これから LPC が大きく影響していることが考えられる。LPC は俺様、少年間では差があるが、優等生は二つの間に位置しているためどちらかに重なりやすいのであろうと考えた。よって LPC に関してはかなり大きな差が出ない限りは特徴量として使用することは難しい。

表 5　最尤法による評価

	俺様	優等生	少年	総合
A〜G(7)	0.4402	0	0.7580	0.4302
A〜F(6)	0.4482	0	0.7581	0.4355
A〜D,F,G(6)	0.4574	0	0.7666	0.4432
A〜C,E〜G(6)	0.5721	0.0027	0.1428	0.4109
A〜D,F(5)	0.4641	0	0.7582	0.4463
A,D〜G(5)	0.4608	0.0050	0.7563	0.4445
A〜C(3)	0.3901	0.6160	0.2345	0.3965

8 おわりに

本研究では、ステレオタイプ性を持つキャラクター音声を話速、F0、LPC を用いて男性キャラクターを約 45% で識別ができることを明らかにした。パラ言語情報に着目し、同様のステレオタイプを持つキャラクターの音声ならば共通する点があると考えた。パラ言語と音響特徴の関連性を調べる従来研究から分析に使用する特徴量を決定した。その決定した特徴量が有効であるかどうかを各特徴量のステレオタイプ別のバタチャリア距離を求めることで評価した。最も距離が遠くなったものは今回提案した特徴量すべてを用いた際の優等生と少年であり、31 ほどの距離があった。最も距離が短かったものと比べるとおよそ 3 倍の差があった。最尤法による評価では話速三種、F0、LPC を用いたものが最も総合精度が高く、0.4463 であった。結果、男性キャラクターの各ステレオタイプを区別する際に音色、F0、話速が有効そうではあるが、特徴量について考え直す必要があることが判明した。

今後の課題として、特徴量に関してはアクセント成分も考慮する必要性があると感じた。また、聴取実験の際の属性の選択肢は対極に位置するキャラクターを集め、特徴が重なりそうなものは外して考えるべきである。自由記述によるアンケートによって選択肢を決定する。そして集まった選択肢からさらにアンケートを取り、厳選する。話速の冒頭や終盤は今回は全体音声長の 2 割と定めたが、ものによってはピークが出ずに上手く求められず、数値が出ていないところが多々ある。閾値の設定方法も考える必要がある。また、音声長の何割かで考えずに単語ごとで考える必要もあると考える。

参考文献

[1] 山岸俊男,”社会心理学”, 新星出版社,2013

[2] 金水敏, ”ヴァーチャル日本語 役割語の謎”, 岩波書店,2003

[3] 穴久保幸,”ポケットモンスター①”,p20

[4] 山本鈴美香,”エースをねらえ！①”,p23

[5] テレビアニメ,”桜蘭高校ホスト部　第一話”,2006

[6] 峯松信明, 広瀬啓吉, 関口真理子, ”話者認識技術を利用した主観的高齢話者の同定とそれに基づく主観的年代の推定”, 情報処理学会論文誌　vol.43, no.7, pp.2186–2196,2002

[7] 石井カルロス寿憲, 石黒浩, 萩田紀博,”韻律および声質を表現した音響特徴と対話音声におけるパラ言語情報の知覚との関連”, 情報処理学会論文誌　vol.47, No.6, pp.1782–pp.1792,2006 J.A.S.A, 50, 2, pp.637-655 ,1971 Proc.IEEE 63, 4, pp.561-580 ,1975 Springer-Verlag, NY, 1976

[8] 山住賢司, 籠宮隆之, 槙洋一, 前川喜久雄,”講演音声の音声的特徴とその印象に対する評価構造モデル”,Japanese Journal of Sensory Evaluation,2007,Vol.11,No.1,30-36