平成 28 年度 卒業論文

男性両声類の女声らしさに関わる特徴量の分析

指導教員 北原 鉄朗 准教授

日本大学文理学部情報科学科

長谷川 翔太

2017 年 2 月 提出

概　　要

近年、男性による女声のような発声 (あるいはその逆) をエンターテイメントの一つとして行う人々が増加し、主に動画投稿サイトなどを通じて**両声類**の名で広く知られるようになった。しかし、両声類の音声の、異性の声らしさに関する研究は進んでいない。異性の声を出すことを目的とした研究に性同一性障害者 (Male to Female transsexuals, MtF) に対する支援に関するものがあるが、両声類はあくまでエンターテイメントを目的としており、自己の本来の声を維持したまま異性の声を出すものであり、目的が異なっている。

本研究では、男性の両声類に焦点を当て、彼らの、成人女性の自然な発声を目的とした女声の音声から抽出した音響特徴量とその音声に対するインターネット上での主観的評価との関係を分析し、主観的評価に影響を与えた音響特徴量について調べる。

音響特徴量を抽出し、複数の分類器で属性選択をした後に分類実験を行ったところ、記憶ベース推論を用いた分類では分類率が 88%という高い結果を出し、この分類で用いられた音響特徴量について、線スペクトル対周波数 (LSP 周波数) の 6 次元の第 1-3 四分位範囲の値が 0.3 以上の音声や LSP 周波数の 5 次元の第 1 四分位数の値が 1.4 以下の音声などは高評価を得られた音声が多い傾向が見られた。また、複数の分類器による属性選択で選ばれた LSP 周波数の 4 次元の第 3 四分位数において、その値が 1.4 前後のものは高評価を得られた音声が多い傾向が見られた。分類実験の結果から、MFCC や LSP 周波数、およびそのΔ特徴量が分類に寄与することが分かった。MFCC や LSP 周波数は声道特性を表すパラメータである

ことから、スペクトルの情報が重要であると考えられるため、特に高い評価を得た話者と低い評価を得た話者の、典型的な音声の/e/の発音の周波数スペクトルに焦点を当てて調べたところ、フォルマント周波数の結果から、両話者ともに/e/の発音であることが想定され、低い評価を得た話者は、高い評価を得た話者より第2~4フォルマントが密集する傾向があった。また、低い評価を得た話者は、高い評価を得た話者に比べてスペクトログラムが疎らであり、フォルマントの時間推移の起伏が激しい傾向があった。MtFの関連研究において、声帯特性に関する記述が多かったため、話者毎の平均基本周波数についても調査したところ、高い評価を得た話者は200Hz~240Hz付近に多く分布し、300Hzを超える話者は成人女性の自然な発声に聞こえるという評価を得にくい傾向があることが分かった。

目　　次

図 目 次

表 目 次

第1章　序　　論

1.1　研究背景

近年、インターネット技術の発達により、人々はインターネット上で自由に会話や動画投稿をするなどといったコミュニケーションを取る機会が増えた。そのような環境の中、男性による女声のような発声 (あるいはその逆) をエンターテイメントの一つとして行う人々が増加し、主に動画投稿サイトなどを通じて**両声類**[1]の名で広く知られるようになった。両声類とは、自分とは異なった性の声を使いこなすことができる人達を指す敬称である。人々に認知され始めた一方、両声類の音声の異性の声らしさに関する研究は進んでいない。異性の声を出すことを目的とした研究に性同一性障害者 (Male to Female transsexuals, MtF) に対する支援に関するものがある [2] が、両声類はあくまでエンターテイメントを目的としており、自己の本来の声を維持したまま異性の声を出すものであり、目的が異なっている。また、両声類が広く認知されるようになったのが比較的近年であることも主な理由であると考えられる。認知されるようになったきっかけの一つとして、ニコニコ動画 [3][4] の存在が挙げられる。ニコニコ動画とは、2006 年にサービスを開始した、コメント機能を用いた動画共有が行えるサイトである。サイト内では動画投稿の他に生放送を行うことやコミュニティを作ることが可能であり、コミュニティには、両声類に関するコミュニティが多数存在する。その中で比較的大きなコミュニティ [5] でも創設は 2008 年と新しい。また、両声類という単語はインターネットスラングであるため、この言葉自体を知る人やその地域が限られてくることも研究が進んでない理由として挙げられる。

1.2 研究の目的

両声類の目指す目標は、多数の人に異性の声に聞こえると評価されることである。多数の人から高い評価を得る両声類の音声と低い評価を得る両声類の音声における特徴の違いは、両声類になりたいと思う人にとって有益な情報であると考えられる。

本研究では、両声類の中でも男性に焦点を当て、男性両声類による女声の音声とそれに対する主観的評価のデータを用意し、音声データから抽出した音響特徴量と主観的評価を用いて、高い評価を得る両声類の音声と低い評価を得る両声類の音声の分類を行うことで、その結果からどの音響特徴量が主観的評価に影響を与えるのかを調べることを目的とする。

1.3 本論文の構成

本論文は次の構成からなる。第2章では、本研究の課題について述べる。第3章では、分類に用いるデータとその特徴抽出について述べる。第4章では、分類実験とその考察について述べる。第5章では、本研究の結論、また課題について述べる。

第2章　本研究へのアプローチ

ここでは、本研究に関連する研究について述べ、本研究との目的の違いについて述べていく。

2.1　関連研究

2.1.1　桜庭らの研究

桜庭らの研究 [2] は、男性から女性へ性転換を希望する性同一性障害者のために音声訓練 (ボイス・セラピー) 法を確立するために、日本文化圏で女声と判定される基本周波数 (F0) の範囲を検討したものである。その結果 80%以上女性と判定された MtF の F0 は 180〜214Hz で、平均 F0 は 193Hz となった。180Hz 以下では男声と判定される率が高くなる一方、平均値以上でも「男子の裏声」と判定される場合があり、高さだけでなく声の質も女声と判定されるためには重要であることが示唆された。

2.1.2　今井田の研究

今井田の研究 [6] は、若い女性の「女性らしさ」に関する言語規範意識の変容によって、音声上にも影響が表れるか調査したものである。女子大学生 25 名によって日本語で読まれた文について声のピッチ測定した結果は、平均基本周波数は 231.54Hz、平均最低値は 201.95Hz、平均最高値は 261.13 であり、日本語を用いる場合と他の言語を用いる場合においてピッチ差があることが観測された。

2.1.3　二村らの研究

二村らの研究 [7] は、男性ホルモン投与中の FTM/GID 23 例を対象に話声位基本周波数を経時的に測定しその変化を解析した。男性ホルモン投与開始時、1ヵ月、2ヵ月、3ヵ月、6ヵ月、9ヵ月、12ヵ月経過時点において楽な発声での母音を録音し基本周波数を測定した。投与開始 1ヵ月経過時点から 3ヵ月経過時点の間に特に急速に音声が低下した。音声の低下は 6ヵ月経過時点まで有意であり、その後 12ヵ月まではわずかな低下があったが有意差は認めなかった。また投与開始時点と 6ヵ月経過時点の話声位基本周波数は有意な相関を認めた。音声が低下すると同時に声帯は発赤および肥大化し、喉頭隆起は顕在化した。男性ホルモン投与により声帯の肥大化と喉頭の枠組みが増大することで音声が低下すると考えられた。

2.2　本研究へのアプローチ

本研究では、MtF の音声の分析ではなく、両声類の音声を分析するというものである。両声類の定義上 [1]、自分の体の性とは異なる性の声を出そうとするという点は、MtF と共通しているといえる。しかし、両声類の多くは、エンターテイメントを求めた、MtF ではない人達であり、異性として生きるために発声する MtF とは、目的や話者の性質が異なる。また、男性両声類には、女声のジャンルとして、幼い女性を意識したロリータ声、成人女性の自然な発声を目的としたナチュラル声、アニメの少女らしい声を意識したアニメ声などが存在する。ナチュラル声は、成人女性の発声を目的としていることから、MtF の発声に近いといえる。しかし、ロリータ声やアニメ声は、MtF の発声の主旨から逸れた、両声類特有の発声であるといえる。今回は、男性両声類のジャンルの中でもナチュラル声に焦点を当てて分析を行う。その際、男性両声類の分析結果と既存の MtF の研究についての比較も行う。

第3章　分類に用いるデータの用意

この章では、本研究で用いた音声データと評価データの詳細と、音響特徴量の抽出について述べる。

3.1　音声データと主観的評価の用意

本研究では、男性両声類の女声らしさに関わる音響特徴量を調べるために、多数の人から高い評価を得る両声類の音声から低い評価を得る両声類の音声まで幅広く用意しなければならない。しかし、両声類として活動する人口の関係上、大勢の両声類とコンタクトを取ることが難しかったため、両声類の音声を直接入手することができなかった。そこで、今回は、ニコニコ動画の両声類コミュニティ[5]でニコニコ生放送[15]を通じて行われた両声類のコンテストの音声と、そのコンテストの結果を分類のデータとして用いることとする。

今回は、両声類のジャンルの中でも「ナチュラル声」に焦点を当てる。男性両声類のナチュラル声の音声を用意し、主観的評価もナチュラル声に聞こえるかどうかということを念頭にそれぞれ評価されたものを用い、分類を行った結果からどの音響特徴量がナチュラル性の主観的評価に影響を与えるのかを調べる。資料として用いたコンテストは2016年4月16日～17日にかけて行われたナチュラル声コンテスト[16]である。

3.1.1　音声データセットの作成

ナチュラル声コンテストの参加者の音声から特に雑音のひどい音声などを除いた 37 人分の両声類の音声データを入手することができた。1 人あたり 10∼20 個程度の音声で、音声データ 1 つあたりの長さは 1∼10 秒程度である。計 534 個の音声データを入手した。

3.1.2　主観的評価について

主観的評価は、話者ごとにナチュラル声コンテストの各視聴者が 1 人 1 票投票できるアンケート形式で採ったものを用いる。アンケートの内容は、

1) ナチュラル声に聞こえる

2) 別の女声に聞こえる

3) 男声に聞こえる

の 3 択であり、ニコニコ生放送のアンケート機能を用いて行われた。選択肢 2) の「 別の女声に聞こえる」とは、アニメ声やロリータ声などといったナチュラル声コンテストの主旨とかけ離れた女声と感じた視聴者が投票する選択肢である。アンケート機能の仕様上、結果は 100 分率で表示される。コンテストの結果は「 ナチュラル声に聞こえる」の割合が上手さの大きな目安になるが、本研究でも「 ナチュラル声に聞こえる」の割合を主観的評価として捉え、各話者について「 ナチュラル声に聞こえる」に投票した視聴者の割合を求め、0∼25%を very low、 25∼50%を low、 50∼75%を high、 75∼100%を very high とした。なお、話者ごとの投票であるため、同一話者に複数の音声がある場合でも評価は一緒である。図 3.1 は各評価に属するデータ数を表す。

表 3.1: 評価ごとの音声データ数

	very low	low	high	very high
データ数	88	144	159	143

3.2 音響特徴量の抽出

表 3.2: emobaseで抽出される特徴量

Feature (26)	Value (19)
PCM intensity	最大値, 最小値
PCM loadness	範囲, 平均
MFCC 1-12	最大/最小の絶対値
LSP Frequency 0-7	線形近似の傾き/切片/誤差
PCM ZeroCrossRate	放物線近似の誤差
voiceProb	標準偏差
F0	分布の歪度/尖度
F0 Envelope	第 k 四分位数 (k=1,2,3)
	第 1-2 四分位範囲
	第 2-3 四分位範囲
	第 1-3 四分位範囲

音響特徴量の抽出には OpenSMILE[17][18] の emobase を用いた。emobase は感情音声やパラ言語情報に関する特徴量を抽出できる。emobase の特徴量セットには 988 個の音響特徴量が含まれている。各音声から表 3.2 に示した特徴量とその時間差分 (Δと呼ぶ) を 10ms 毎に抽出し、それぞれについて時間軸方向に表 3.2 の Value に示した値を求める。

intensity は、音の物理的な刺激の程度を表す。loadness は、人が感じる音の大きさを表す。MFCC は、メル周波数ケプストラム係数 (Mel Frequency Cepstrum

Coefficients) を表す声道特性を表すパラメータである。LSP 周波数は、線スペクトル対 (line spectral pairs) を意味する、線形予測係数を表現するために使われる声道特性を表すパラメータである。zcr (zero crossing rate) は、音の波形データが 0 を交差する頻度を表すパラメータであり、この値が大きくなるほど騒がしい音声であると捉えられる。voiceProb (voicing probability) は、自己相関係数 (ACF : autocorrelation function) から導き出されたパラメータを表す。F0 は、基本周波数を意味する、音高を司る音響特徴量である。F0 envelope は、F0 の包絡を表すパラメータである。

第4章　分類実験

今回の実験では、複数の分類器を用いて音響特徴量と主観的評価の属性選択と分類を行った。属性選択において共通して選ばれた音響特徴量や分類精度の高かった結果などについて考察する。

4.1　実験条件

分類で用いるデータセットは、3章で用意した、抽出された音響特徴量と主観的評価を元に付けられたランクを用いる。分類にはWeka[19]を用いた。分類を行う前に属性選択を行い、988個の音響特徴量の中から識別に有用なものだけを選び出す。属性選択には、サンプリングした訓練データに対し実際に分類アルゴリズムを適用し、分類精度が最高となる部分集合を得る、ラッパーアプローチと呼ばれる手法を用いた。属性選択で用いた分類器は、ベイジアンネットワーク (BayesNet)、単純ベイズ (NaiveBayes)、RBFネットワーク (RBFNetwork)、記憶ベース推論 (IBk)、決定木 (J48) であり、分類の際も同様の分類器を用いた。また、分類の手法は10-fold cross validationである。複数の分類器を用いた理由として、属性選択の際に選ばれた音響特徴量を比較するためである。

表 4.1: 属性選択と 分類の結果

分類器	分類率	適合率	再現率
BayesNet(18)	0.618	0.617	0.618
NaiveBayes(26)	0.742	0.741	0.742
RBFNetwork(15)	0.702	0.701	0.702
IBk(30)	0.880	0.882	0.880
J48(14)	0.642	0.644	0.642

表 4.2: BayesNet の分類結果

		Prediction				
		very low	low	high	very high	recall
Actual	very low	37	18	18	15	0.42
	low	3	96	36	9	0.667
	high	15	24	93	27	0.585
	very high	8	7	24	104	0.727
precision		0.587	0.662	0.554	0.671	

4.2　実験結果

4.2.1　分類結果

表 4.1 は、分類実験の結果を 分類器ごとに示している。また、分類器名の後の () の中の数は、属性選択によって選ばれ、分類の際に用いられた属性の数を 表している。表 4.2∼4.6 は、それぞれの分類結果の詳細である。表 4.1 の分類結果の中でも IBk による 分類は、分類率 88%、適合率 88.2%、再現率 88%であり 、分類の精度が高いという 結果となった。IBk は、局所的な推測に有利な 方法で、外れ値の影響を 受けにくいという 長所がある [20]。分類に用いた音響特徴量において、高い評価を 得た 音声データが多く 分布する 帯域や低い評価を 得たデータが多く 分布する

表 4.3: NaiveBayes の分類結果

		Prediction				
		very low	low	high	very high	recall
Actual	very low	62	10	11	5	0.705
	low	5	111	21	7	0.771
	high	13	24	104	18	0.654
	very high	7	9	8	119	0.832
precision		0.713	0.721	0.722	0.799	

表 4.4: RBFNetwork の分類結果

		Prediction				
		very low	low	high	very high	recall
Actual	very low	53	10	16	9	0.602
	low	9	103	20	12	0.715
	high	14	19	107	19	0.673
	very high	5	9	17	112	0.783
precision		0.654	0.73	0.669	0.737	

表 4.5: IBk の分類結果

		Prediction				
		very low	low	high	very high	recall
Actual	very low	80	2	5	1	0.909
	low	6	133	3	2	0.924
	high	5	11	132	11	0.83
	very high	6	7	5	125	0.874
precision		0.825	0.869	0.91	0.899	

表 4.6: J48 の分類結果

		Prediction				
		very low	low	high	very high	recall
Actual	very low	58	15	10	5	0.659
	low	18	80	30	16	0.556
	high	19	26	97	17	0.61
	very high	10	13	12	108	0.755
precision		0.552	0.597	0.651	0.74	

帯域があるのではないかと考えられる。表 4.7~4.11 は、属性選択で選ばれ、各分類で用いられた音響特徴量である。MFCC や LSP 周波数、およびその Δ 特徴量が多く 選ばれていることが分かる。MFCC や LSP 周波数は声道特性を 表すパラメータであるため、スペクト ルの情報が重要であると考えられる。

4.2.2　精度の高かった分類で用いられた音響特徴量

今回の分類で最も精度の高かった IBk の分類の際に用いられた 30 種類の音響特徴量において、属性毎に評価値を算出して閾値以上の属性を 残すフィルターアプローチを 行いランク付けをして、それら の音響特徴量の中でも 特に識別に有用なものを 調査した。

図 4.1 は IBk の分類の際に用いられた 30 種類の音響特徴量において特に重みの大きかった 2 つの特徴量の散布図である。また、図 4.1 の下の図は、その中でも very high と very low がラベル付けされた音声のみの散布図である。図 4.1 から LSP 周波数の 6 次元の第 1-3 四分位範囲の値が 0.3 以上の音声は高い評価を 得たものが多く、また、LSP 周波数の 5 次元の第 1 四分位数の値が 1.4 以下の音声も 高い評価を得たものが多い。しかし、LSP 周波数の 6 次元の第 1-3 四分位範囲の値が 0.3 以下かつ LSP 周波数の 5 次元の第 1 四分位数の値が 1.4 以上の音声は高い評価の音声

表 4.7: BayesNet で用いられた音響特徴量

features
MFCC 2 次元 平均/第 1 四分位数
MFCC 4 次元 第 2 四分位数
MFCC 9 次元 第 1 四分位数
MFCC 10 次元 最大値
MFCC 11 次元 第 1 四分位数
MFCC 12 次元 第 1 四分位数
LSP Frequency 1 次元 範囲
LSP Frequency 2 次元 分布の歪度
LSP Frequency 3 次元 分布の尖度
LSP Frequency 4 次元 第 3 四分位数
F0 envelope 第 2 四分位数
Δ PCM intensity 線形近似の傾き
Δ PCM loudness 最大値
Δ MFCC 2 次元 範囲/線形近似の誤差
Δ MFCC 7 次元 第 1-2 四分位範囲
Δ LSP Frequency 6 次元 標準偏差

表 4.8: NaiveBayes で用いられた音響特徴量

features
PCM loadness 標準偏差
MFCC 1 次元 分布の歪度
MFCC 3 次元 第 3 四分位数
MFCC 4 次元 分布の尖度
MFCC 6 次元 標準偏差/第 3 四分位数
MFCC 10 次元 最大値/標準偏差
MFCC 11 次元 第 1 四分位数
LSP Frequency 0 次元 第 2-3 四分位範囲
LSP Frequency 2 次元 範囲/第 1 四分位数/第 1-2 四分位範囲
LSP Frequency 3 次元 最大値/平均
LSP Frequency 4 次元 第 3 四分位数
LSP Frequency 5 次元 標準偏差
LSP Frequency 6 次元 標準偏差
voiceProb 第 1-2 四分位範囲
Δ MFCC 4 次元 標準偏差
Δ MFCC 6 次元 第 2 四分位数
Δ MFCC 11 次元 最小値
Δ MFCC 12 次元 分布の歪度
Δ LSP Frequency 1 次元 線形近似の誤差
Δ LSP Frequency 4 次元 標準偏差
Δ PCM zcr 第 3 四分位数

表 4.9: RBFNetwork で用いられた音響特徴量

features
PCM loadness 線形近似の誤差
MFCC 2 次元 最大値
MFCC 3 次元 第 1 四分位数/第 2-3 四分位範囲
MFCC 4 次元 第 1 四分位数
MFCC 7 次元 第 3 四分位数
MFCC 8 次元 平均
MFCC 11 次元 第 2 四分位数
LSP Frequency 2 次元 第 1 四分位数
LSP Frequency 4 次元 第 3 四分位数
Δ MFCC 2 次元 第 2-3 四分位範囲
Δ MFCC 9 次元 分布の尖度
Δ LSP Frequency 3 次元 第 2 四分位数
Δ LSP Frequency 4 次元 放物線近似の誤差
Δ F0 第 1 四分位数

表 4.10: IBk で用いられた音響特徴量

features
PCM intensity 放物線近似の誤差/標準偏差
MFCC 2 次元 第 1 四分位数/第 3 四分位数
MFCC 3 次元 第 1 四分位数
MFCC 4 次元 放物線近似の誤差
MFCC 6 次元 平均/標準偏差
MFCC 10 次元 第 3 四分位数
MFCC 11 次元 平均
MFCC 12 次元 第 3 四分位数
LSP Frequency 2 次元 最大値
LSP Frequency 3 次元 分布の尖度/第 1 四分位数
LSP Frequency 5 次元 最大値/第 1 四分位数
LSP Frequency 6 次元 第 2 四分位数/第 1-3 四分位範囲
PCM zcr 第 1-2 四分位範囲
F0 第 2 四分位数
Δ PCM intensity 平均/第 3 四分位数
Δ MFCC 1 次元 線形近似の傾き
Δ MFCC 2 次元 平均
Δ MFCC 4 次元 線形近似の傾き
Δ MFCC 5 次元 放物線近似の誤差
Δ MFCC 6 次元 平均
Δ MFCC 9 次元 線形近似の誤差
Δ MFCC 10 次元 線形近似の傾き
Δ LSP Frequency 3 次元 線形近似の傾き

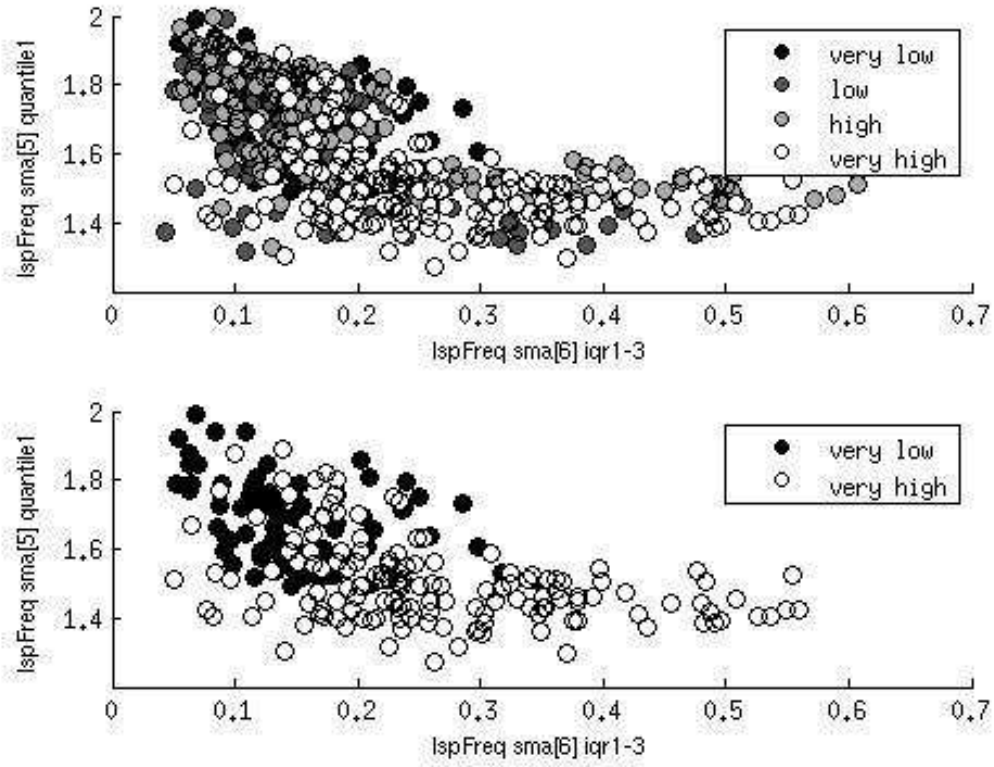

図 4.1: 識別に有用だった特徴量の分布

表 4.11: J48 で用いられた音響特徴量

features
MFCC 2 次元 範囲/平均
MFCC 10 次元 最小値/範囲/第 3 四分位数
LSP Frequency 1 次元 放物線近似の誤差
LSP Frequency 3 次元 第 1 四分位数
LSP Frequency 4 次元 第 3 四分位数
LSP Frequency 5 次元 第 3 四分位数
LSP Frequency 6 次元 線形近似の切片
Δ PCM loadness 第 1 四分位数
Δ LSP Frequency 5 次元 標準偏差/第 1-2 四分位範囲
Δ LSP Frequency 7 次元 標準偏差

と低い評価の音声が混じり合っている。

表 4.12: 上位 2 つの音響特徴量の分類結果

	分類率	適合率	再現率
高評価 (very high) の音声	0.741	0.757	0.741
低評価 (very low) の音声	0.614	0.593	0.614

表 4.12 は、特に重みの大きかった 2 つの音響特徴量だけで分類を行った結果である。高評価だった音声は正確に分類された割合が高く、低評価だった音声は正確に分類された割合が少なかった。これは、高評価のみが分布している範囲は広いが、低評価が分布する範囲には高評価のデータが多数分布していることが原因であると考えられる。

表 4.13: 複数の分類で用いられた音響特徴量

features	number of times the selected
MFCC 2 次元 平均	2
MFCC 2 次元 第 1 四分位数	2
MFCC 3 次元 第 1 四分位数	2
MFCC 6 次元 標準偏差	2
MFCC 10 次元 最大値	2
MFCC 10 次元 第 3 四分位数	2
MFCC 11 次元 第 1 四分位数	2
LSP Frequency 2 次元 第 1 四分位数	2
LSP Frequency 3 次元 分布の尖度	2
LSP Frequency 3 次元 第 1 四分位数	2
LSP Frequency 4 次元 第 3 四分位数	4

4.2.3 複数の分類で用いられた音響特徴量

表 4.13 は、属性選択で選ばれた選ばれた音響特徴量について、複数の分類で共通して選ばれ、用いられた音響特徴量である。中でも、LSP 周波数の 4 次元の第 3 四分位数は 4 種の分類で用いられ、識別に有用な特徴量であったことが伺える。

図 4.2 は、属性選択で最も多く共通して選ばれ、分類に用いられた LSP 周波数の 4 次元の第 3 四分位数の、高い評価 (very high) を得た音声と低い評価 (very low) を得た音声のヒストグラムである。この図から、LSP 周波数の 4 次元の第 3 四分位数の値が 1.4 前後の音声は高い評価を得た音声の割合が高く、値が 1.4 から離れた音声は低い評価を得た音声の割合が高いということが分かり、識別の一端を担う特徴量であるといえる。

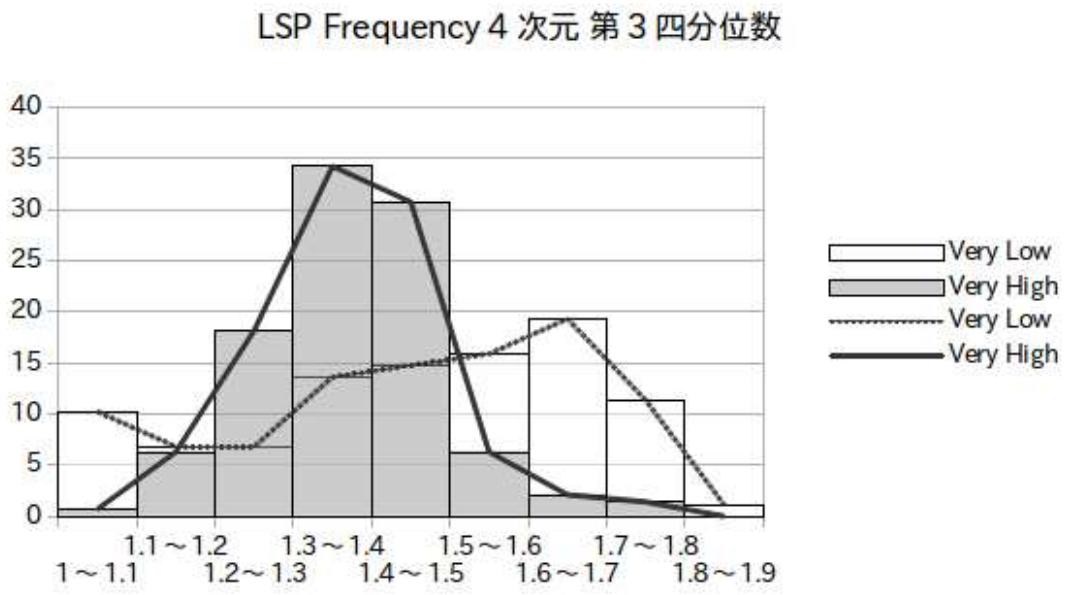

図 4.2: 識別に有用だった特徴量の分布2

4.3　周波数スペクトル

分類実験の結果、MFCC や LSP 周波数に関する音響特徴量が分類に寄与することが分かった。MFCC や LSP 周波数は声道特性を表すパラメータであるため、スペクトルの情報が重要であると考えられる。そこで、分類に用いた音声データの中で高い評価 (very high) を得た音声と低い評価 (very low) を得た音声について、それぞれの典型的な音声の周波数スペクトルについて調べた。分類に用いた音声データは、話者ごとに内容の違う会話文であったため、それらの音声の比較的切り取りやすかった/e/の発音を対象とした。/e/の発音の両端は前後の発話による影響を受けると考えられるため、定常部分のみを対象とした。また、調査には WaveSurfer[21] を用いた。

図 4.3 と図 4.4 は、それぞれ高い評価を得た話者と低い評価を得た話者の/e/の発音のスペクトログラムである。図 4.5 と図 4.6 は、それぞれ高い評価を得た話者

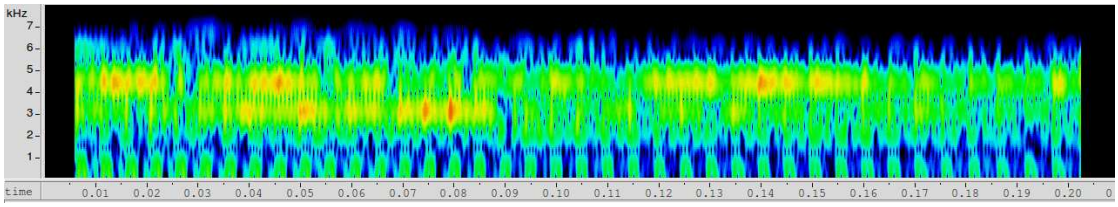

図 4.3: 高い評価を得た話者の/e/のスペクトログラム

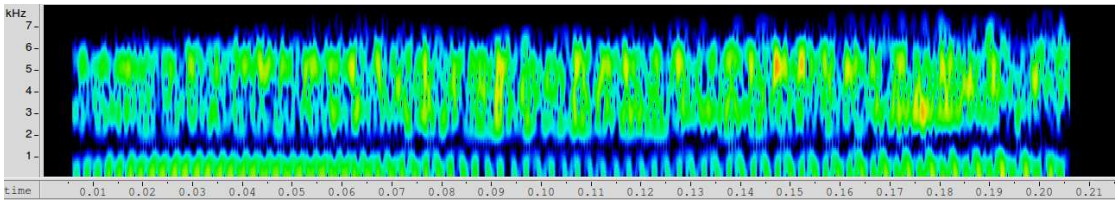

図 4.4: 低い評価を得た話者の/e/のスペクトログラム

と低い評価を得た話者の/e/の発音のフォルマントの時間推移であり、赤・青・黄・緑の 4 本の横線は、周波数軸の低い方から第 1~4 フォルマントを表す。図 4.3 と図 4.4 を比較すると、低い評価を得た話者は、高い評価を得た話者と比較すると、全体的に時間軸方向の変化が激しいことが分かる。図 4.5 と図 4.6 を比較すると、低い評価を得た話者は、高い評価を得た話者と比較すると全体的に第 2~4 フォルマントが密集しており、フォルマントの時間推移の起伏が激しい傾向があった。

高い評価を得た話者と低い評価を得た話者の/e/の発音の第 1~4 フォルマントの時間平均の値を表 4.14 に、標準偏差の値を表 4.15 に表す。表 4.14 の結果から、両話者の発音は/e/であったと想定される。また、低い評価を得た話者のフォルマントの時間平均は、高い評価を得た話者に比べて第 2~3 フォルマント間の差が 362Hz、第 3~4 フォルマント間の差が 460Hz 小さく、第 2~4 フォルマント間の差は 822Hz 小さいことが分かり、密集していたことが分かった。また、表 4.15 の結果から、低い評価を得た話者の標準偏差は、高い評価を得た話者に比べて第 1~4 フォルマントすべてにおいてその値が大きいことが分かり、フォルマントの起伏が激しいこ

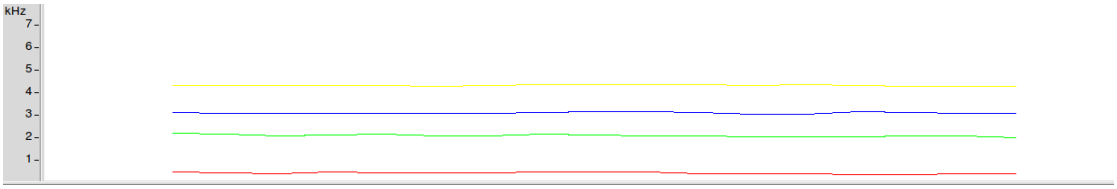

図 4.5: 高い評価を 得た話者の/e/のフォルマント の時間推移

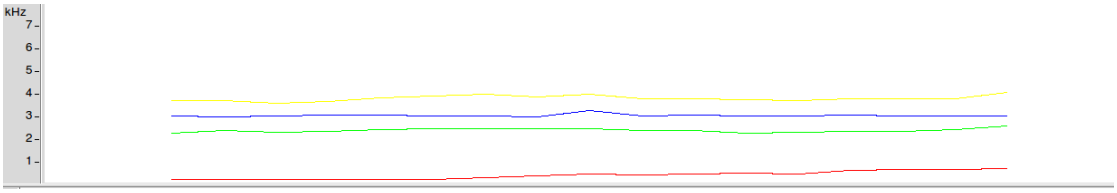

図 4.6: 低い評価を 得た話者の/e/のフォルマント の時間推移

とが分かった。実際に 2 つの音声を 聞いたところ、低い評価を 得た話者は、高い評価を 得た話者に比べ、声が裏返り 震えたように聞こえ、発声が安定していない印象を 受けた。

4.4　基本周波数

既存の MtF に関する研究 [2] では、「 80%以上女性と 判定された MtF の平均基本周波数は 193Hz であった」という 研究結果があり、声の高さも 女声らしさに重要な情報であると 予想されるが、今回の分類では、基本周波数に関する特徴量が識別に用いられることは少なかったため、話者毎の平均基本周波数に焦点を 当て、その分布を 調べた。特徴抽出には、音声分析変換合成システムの WORLD[22] を 用いた。

図 4.7 は、37 人の話者毎の平均基本周波数である。横軸の rating とは、ナチュラル声コンテスト の評価において、「 ナチュラル声に聞こえる」と答えた割合である。この図 4.7 から、平均基本周波数が 200Hz∼240Hz 付近では、低い評価を 得た

表 4.14: /e/のフォルマント 周波数の時間平均

	F1(Hz)	F2(Hz)	F3(Hz)	F4(Hz)
高評価 (very high) の音声	429	2098	3100	4329
低評価 (very low) の音声	436	2414	3054	3823

表 4.15: /e/のフォルマント 周波数の標準偏差

	F1	F2	F3	F4
高評価 (very high) の音声	36	42	35	23
低評価 (very low) の音声	174	74	58	118

話者から高い評価を得た話者まで幅広く分布しているが、高い評価を得た話者はこの付近に集まる傾向があり、分布に収束性があることが分かる。女性の声の高さに関する研究 [6] によると、女子大学生 25 名が日本語文を読んだときの平均基本周波数は 231Hz、その範囲が 201∼261Hz という結果であり、ナチュラル声に聞こえるという評価を多く受けた両声類の音声の分布とほぼ一致する結果となった。MtF の関連研究 [2] と比較すると、平均基本周波数は両声類の方が全体的に高い傾向が見られた。また、平均基本周波数が 300Hz を超える話者については、ナチュラル声に聞こえるという評価を得にくい傾向があることが分かる。MtF に関する調査 [11] でも、高すぎる基本周波数は女性の声と判別されにくい傾向があり、MtF の音声と両声類の音声で一致する点であったといえる。図 4.7 の分布の相関関係を調べたところ、相関係数が-0.0801、p 値が 0.6376 であり、相関関係があるとはいえない結果となった。故に、ナチュラル声に聞こえるという評価を得られる声の高さの帯域は存在するが、声の高さだけでは高い評価を得ることができないといえる。

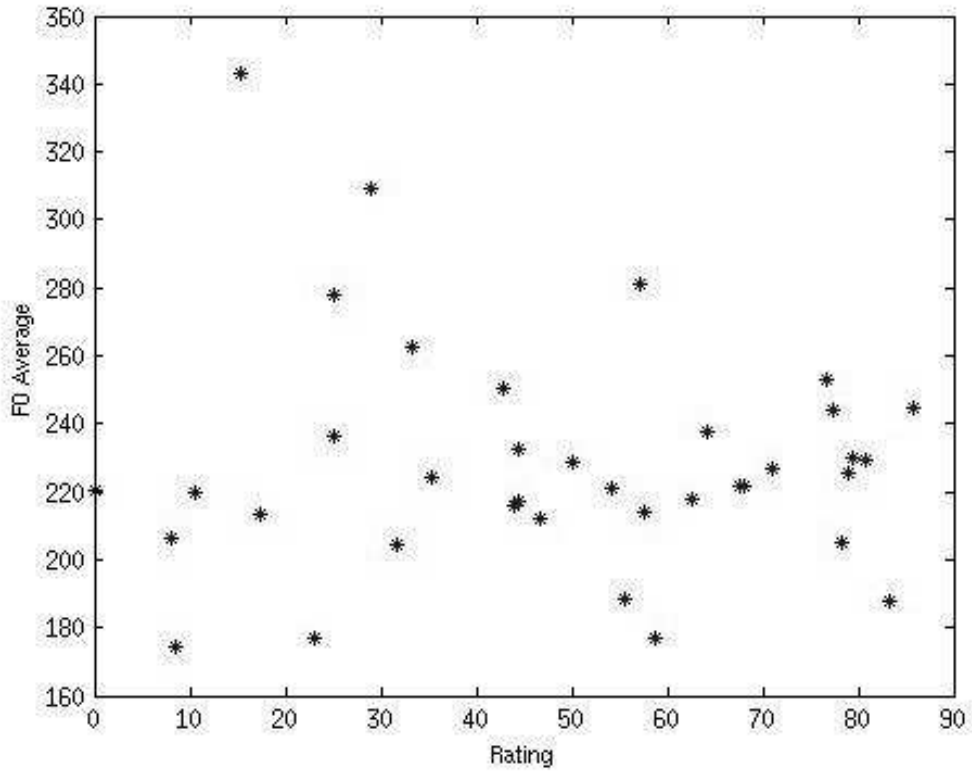

図 4.7: 話者毎の平均基本周波数の分布

第5章　結　　論

本研究では、男性両声類の女声に焦点を当て、彼らの、成人女性の自然な発声を目的とした音声から抽出した音響特徴量とその音声に対する主観的評価を用いて、高い評価を得る両声類の音声と低い評価を得る両声類の音声の分類を行うことで、その結果からどの音響特徴量が両声類の音声のナチュラル性の主観的評価に影響を与えているのかを調べることを目的とした。

分類実験の結果、IBk を用いた分類では分類率が 88%という高い結果であり、MFCC や LSP 周波数、およびそのΔ特徴量が分類に寄与することが分かった。音響特徴量についても、LSP 周波数の 6 次元の第 1-3 四分位範囲の値が 0.3 以上の音声や LSP 周波数の 5 次元の第 1 四分位数の値が 1.4 以下の音声などは高評価を得られた音声が多く分布していた。また、LSP 周波数の 4 次元の第 3 四分位数において、その値が 1.4 前後のものは高評価を得られた音声が多い傾向が見られ、ナチュラル声コンテストの評価に関わった音響特徴量であるといえる。

MFCC や LSP 周波数は声道特性を表すパラメータであることから、スペクトルの情報が重要であると考えられるため、高い評価を得た話者と低い評価を得た話者の、典型的な音声の/e/の発音の周波数スペクトルに焦点を当てて調べた。その結果、低い評価を得た話者は、高い評価を得た話者と比較すると、全体的にスペクトログラムの時間軸方向の変化が激しいことが分かった。フォルマント周波数の結果からは、両話者ともに/e/と発音していることが想定され、低い評価を得た話者のフォルマントの時間平均は、高い評価を得た話者に比べて第 2~3 フォルマント間の差が 362Hz、第 3~4 フォルマント間の差が 460Hz 小さく、第 2~4 フォル

マント間の差は822Hz小さいことが分かり、密集していたことが分かった。また、低い評価を得た話者の標準偏差は、高い評価を得た話者に比べて第1~4フォルマントすべてにおいてその値が大きいことが分かり、フォルマントの起伏が激しいことが分かった。

MtFの既存研究では、平均基本周波数に関する研究結果があったため、話者ごとの声の高さも重要であると考え、調査を行った。その結果、高い評価を得た話者は200Hz~240Hz付近に多く分布し、MtFの関連研究と比較すると、全体的にその値が高い傾向が見られた。平均基本周波数が300Hzを超える話者はナチュラル声に聞こえるという評価を得にくい傾向があることが分かったが、この点に関してはMtFの関連研究と一致する傾向であった。平均基本周波数と主観的評価の相関関係を調べたところ、相関係数が-0.0801、p値が0.6376であり、相関関係があるとはいえない結果となった。以上の結果から、ナチュラル声に聞こえるという評価を得られる声の高さの帯域が存在するが、声の高さだけではなく、声道特性も評価において必要な情報であるといえる。

今後の課題としては、高評価を得た音声と低評価を得た音声における、今回比較しなかった他の音響特徴量についても比較を行い、更なる識別に有効な音響特徴量について調べていきたい。また、両声類のアニメ声やロリータ声などといった、今回分析できなかった両声類のジャンルについても調査したい。

参考文献

[1] ニコニコ大百科 両声類
http://dic.nicovideo.jp/a/%E4%B8%A1%E5%A3%B0%E9%A1%9E

[2] 櫻庭 京子他: “女性と判定された性同一性障害者 (MtF) の声の基本周波数”、電子情報通信学会技術研究報告. SP, 音声 102 (749), 49-52, 2003-03-20

[3] niconico
http://www.nicovideo.jp/

[4] ニコニコ大百科 ニコニコ動画
http://dic.nicovideo.jp/a/%E3%83%8B%E3%82%B3%E3%83%8B%E3%82%B3%E5%8B%95%E7%94%BB

[5] niconico 両声類コミュニティ
http://com.nicovideo.jp/community/co4805

[6] 今井田 亜弓: “若い日本人女性のピッチ変化に見る文化的規範の影響”、言語文化論集 27 (2), 13-26, 2006

[7] 二宮 吉継他: “性同一性障害 Female to Male 症例の男性ホルモン投与による話声位基本周波数の継時的変化”、音声言語医学 56 (4), 348-356, 2015

[8] 櫻庭 京子他: “話者認識技術を用いた性同一性症者 (MtF) の音声に対する男声度・女声度の自動推定とその臨床応用”、電子情報通信学会技術研究報告. SP, 音声 105 (686), 29-34, 2006-03-21

[9] 櫻庭 京子他: “男性から女性への性別の移行を希望する性同一性障害者 (MtF) の発話音声の分類に関する試案”、電子情報通信学会技術研究報告. SP, 音声 106 (613), 1-5, 2007-03-19

[10] 櫻庭 京子他: “女性と判定される声の特徴 : 性同一性障害者の話声位”、音声言語医学 50 (1), 14-20, 2009-01-20

[11]「 女性と判断される声の特徴について」早田 直著
http://wasadasan.com/repo2/koepass.pdf

[12] 管村 昇他: “線形予測係数の線スペクトル表現とその統計的性質”、電子情報通信学会論文誌 A, Vol. J64-A, No. 4, 323-330, 1981-04-25

[13] 森勢将雅，河原英紀，西浦敬信: “基本波検出に基づく高 SNR の音声を対象とした高速な F0 推定法” 電子情報通信学会 論文誌 D，vol. J93-D, no. 2, pp. 109-117, Feb. 2010

[14] M. Morise, F. Yokomori, and K. Ozawa, “WORLD: a vocoder-based high-quality speech synthesis system for real-time applications,” IEICE transactions on information and systems, vol. E99-D, no. 7, pp. 1877-1884, 2016

[15] ニコニコ生放送
http://live.nicovideo.jp/

[16] 第 4 回ナチュラル声コンテスト
http://live.nicovideo.jp/watch/lv259426522

[17] Florian Eyben, Felix Weninger, Florian Gross, Bjorn Schuller: “Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor”, In Proc. ACM Multimedia (MM), Barcelona, Spain, ACM, ISBN 978-1-4503-2404-5, pp. 835-838, October 2013.

[18] Florian Eyben, Martin Wollmer, Bjorn Schuller: “openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor”, In Proc. ACM Multimadia (MM), ACM, Florence, Italy, ACM, ISBN 978-1-60558-933-6, pp. 1459-1462, October 2010.

[19] Weka
http://www.weka-jp.info/index.php/weka-jp/2011-05-25-10-58-08

[20]「 データマイニング手法」マイケル J. A. ベリー・ゴートン・リノフ著　海文堂　1999

[21] WaveSurfer
https://sourceforge.net/projects/wavesurfer/

[22] WORLD
http://ml.cs.yamanashi.ac.jp/world/index.html

謝　　辞

本研究を進めるにあたり、北原鉄朗准教授から丁寧かつ熱心なご指導を賜りました。また、研究に協力してくださった大野涼平さんに感謝致します。