

修 士 論 文

話者認識技術に基づく知覚的女声度の
自動推定

平成19年2月2日 提出

指導教員 峯松 信明 助教授

情報理工学系研究科 電子情報学専攻

56438 丸山 和孝

概要

性同一性障害者の中で、特に MtF（男性から女性へ性別の移行を希望するもの）を対象として、音声がどの程度女性らしく聞こえるかを、コンピュータにより自動的に推定する手法を提案した。声の性別に関しては、従来は母音ごとのフォルマントを分析するなどの方法がとられていたため、受容できる音声に限られていた。本論分で提案した手法を用いれば、テキスト非依存の話者照合技術を応用することで、発話内容を問わず、連続的な音声に対して安定的に音響的女声度を評価することができる。

また、聴取実験から、第三者にとって MtF の音声に対してどの程度女声らしく聞こえるかをラベリングした。ここでは、人間の判断する声の女性らしさ（知覚的女声度）を定義し、MtF の音声に対するラベルとした。そして、コンピュータにより算出された音響的女声度と、聴取実験によって得られた知覚的女声度の比較から、より女性らしく聞こえる声で話すためには、ただ声を高くするだけでなく、声道形状を適切に制御することも必要であるという知見を得た。

次に、線形回帰分析により、聴取実験によって得られる知覚的女声度を予測した。知覚的女声度の予測値と実際の聴取結果の相関係数は、0.86 となり、良好な結果を得られた。この相関係数と対応する人間同士の評価の相関係数の平均が 0.88 であったことから、本論分で構築した知覚的女声度自動推定器はひとりの人間とみなせる程度の性能を持っていることがわかった。この推定器にはインターフェースを実装し、臨床の場で使用していただき、好評を得ている。

目次

第1章	序論	1
1.1	はじめに	1
1.2	本論文の構成	2
第2章	性同一性障害	3
2.1	性同一性障害	3
2.1.1	性同一性障害の定義	3
2.1.2	同性愛などとの混同について	4
2.1.3	日本における性同一性障害	4
2.2	MtF と声の問題	5
2.3	声帯手術による声の女性化の試みとその問題点	6
2.4	トランスセクシュアルボイスセラピー	6
2.4.1	声の女性らしさ決める要因	6
2.4.2	声の高さ	7
2.4.3	声道形状の制御	7
2.4.4	女性らしい話し方	7
2.4.5	女声を判断する評価者の問題	8
第3章	音声生成の原理と工学的モデル	9
3.1	ソースフィルタモデル	9
3.2	音響特徴量	9
3.2.1	分節的特徴	9
3.2.2	韻律的特徴	10
第4章	音声に含まれる情報と声の女声らしさと先行研究	12
4.1	音声に含まれる情報	12
4.2	声の女性らしさ・性別に関する先行研究	12
4.2.1	声道模型を用いた合成音声による研究	12
4.2.2	スペクトルの“山”を用いた声の男女の研究	13
4.2.3	男声連続音声の女声への変換する実験	14
4.2.4	声道面積関数をパラメータとし主成分分析を用いる手法による研究	15
4.2.5	声の性別の知覚に関する先行研究	15
4.2.6	性別認識に関する先行研究	17
4.3	声の女性らしさに関する先行研究のまとめ	18

第 5 章	話者認識技術	20
5.1	話者認識	20
5.2	話者認識の種類	20
5.3	テキスト依存性	20
5.4	認識手法	21
5.4.1	GMM ベースの話者照合	21
5.4.2	その他の手法を用いた話者認識	22
5.5	性能の評価	22
5.6	話者照合を応用した先行研究	22
5.6.1	年齢推定	22
第 6 章	知覚的女声度の聴取実験	25
6.1	性同一性障害者の音声の収録	25
6.2	知覚的女声度の定義と聴取実験	25
6.2.1	実験条件	25
6.2.2	知覚的女声度の分布	26
6.3	評価者間の相関と評価者内の相関	26
6.4	2 値判定による予備実験	27
第 7 章	声の女性らしさ推定の枠組み	28
7.1	既存の技術に基づく孤立母音に対する女声度推定と問題点	28
7.2	話者照合技術に基づく声の女性らしさの定義	28
7.3	女性らしさの程度を算出する妥当性	29
7.4	音響特徴量の検討	29
7.4.1	ボイスセラピーとの関連からの検討	29
7.4.2	使用する音響特徴量	30
7.5	GMM ベースのモデルの構築	30
7.6	線形回帰による拡張と予測値の算出	31
7.7	実験に用いたコーパス	31
7.7.1	新聞記事読み上げ音声コーパス (JNAS)	32
第 8 章	実験結果	33
8.1	単純な推定の結果	33
8.1.1	議論	33
8.2	線形回帰による予測値の算出	33
8.2.1	線形回帰係数	35
8.3	知覚的女声度のラベルとして 2 値の聴取結果を用いた場合	36
8.3.1	予測値の算出	36
8.3.2	GID 話者の音声によるモデル	36
第 9 章	女声度推定器のインターフェース化と臨床の場面への導入	40
9.1	インターフェースの実装	40
9.2	臨床応用	40

目 次

3.1	ソースフィルタ音声生成モデル	10
3.2	f0 パターンの例	11
4.1	声道模型	13
4.2	声道長と基本周波数のマッチング	14
4.3	音源波形	14
4.4	各母音のスペクトル形状	15
4.5	各母音の声道断面積関数	16
4.6	声道断面積関数の主成分分析の第 1 軸と第 2 軸に対する各母音と男女の広がり	17
4.7	声の性別の連続的变化に対する知覚実験結果	18
4.8	3 カテゴリーでの知覚実験結果	19
4.9	テンプレート作成手順	19
5.1	話者識別	21
5.2	話者照合	21
5.3	GMM ベース話者照合の概略図	22
5.4	知覚的年齢の分布	23
5.5	知覚的年齢とその推定値の相関	24
6.1	知覚的女声度のヒストグラム	26
7.1	システムの概観	32
8.1	声道特性に基づく音響的女声度と聴取結果による知覚的女声度の関係	34
8.2	音源特性に基づく音響的女声度と聴取結果による知覚的女声度の関係	34
8.3	知覚的女声度の予測値と知覚的女声度の関係 (全評価者平均)	38
8.4	知覚的女声度の予測値と知覚的女声度の関係 (男性評価者平均)	38
8.5	知覚的女声度の予測値と知覚的女声度関係 (女性評価者平均)	38
8.6	MFCC を音響特徴量として用いた場合の音響的女声度と 2 値評価による聴取結果の相関	39
8.7	$\log F_0$ を音響特徴量として用いた場合の音響的女声度と 2 値評価による聴取結果の相関	39
8.8	線形回帰による予測と 2 値評価による聴取結果の相関	39
9.1	知覚的女声度推定器のインターフェース	41

9.2	知覚的女声度推定器の実際の使用の様子	41
9.3	知覚的女声度推定器の実際の使用における画面表示	42

表 目 次

4.1	音声に含まれる情報	12
6.1	話者間・話者内の評価値の相関	27
6.2	ある評価者とそのほかの評価者の平均の相関	27
7.1	分析条件	31
7.2	新聞記事読み上げ音声コーパスの概要 [34]	32
8.1	音響的女声度と知覚的女声度の相関	35
8.2	知覚的女声度とその予測値の相関	35
8.3	発話オープンな条件下での知覚的女声度とその予測値の相関	35
8.4	線形回帰係数	36

第1章 序論

1.1 はじめに

近年の音声工学における技術の発展と、最近の急速な計算機能力の向上により、音声認識の精度は格段の向上を遂げている。一方、それに伴い、言語情報以外の情報の認識、たとえば感情の認識も活発に行われるようになってきている。

ところで、近年になり性同一性障害という言葉がテレビ等のメディアでもよく耳にするようになった。テレビドラマで性同一性障害の悩みを抱える登場人物が主題となったり、スポーツの選手が性同一性障害であることを公表したといったことは記憶に新しい。性同一性障害とは簡単に言えば、もって生まれた生物学的な性と自己が認識する性が一致しないことによって引き起こされる障害である。

この障害をもつ人々が直面している問題はさまざまである。性同一性障害という事象自体が、世間に知られるようになってまだ間もないこともあり、性同一性障害者に対する差別や偏見は少なくない。また、性同一性障害者、たとえば MtF (Male to Female、男性から女性へ性別の移行を希望するもの) であれば肉体などの女性化を希望する。容姿・外見に関する問題はよく知られているが、これは性転換手術(性別適合手術ともよばれる)やホルモン投与により解決される。実際には、完全な女性化を達成する上で、一番の問題となるのは声である。とくに MtF に関しては、ホルモン治療や声帯手術などによる声の女性化は難しく、性別移行前の性が男性であると知られずに女性として生活することを望む MtF が、生物学的には男性であると見抜かれる第一の原因が声である。

MtF の声を女性化する方法はまだ確立されていないが、現在のところトランスセクシュアルボイスセラピー(Transsexual Voice Therapy)が最も効果的といわれている。ボイスセラピーでは、セラピストが MtF に対し、女性らしい声の出し方の訓練をする。いずれにせよ、声の女性化が達成されたかどうかを知るためには、第三者による聴取実験により MtF の声がどの程度女性の声として判定されるかを調べる必要がある。しかし、この聴取実験は実験の意図を知ることによる先入観の影響を大きく受けてしまうため、聴取者の確保が非常に大きな労力となってしまう。

本研究ではこの問題を解決するシステムの構築を主眼とする。まず、コンピュータによる計算で、音響的な声の女性らしさ(音響的女声度)を求めることを試みる。既存の技術としては、主に母音ごとにフォルマント周波数を分析し、その男女の違いから女性らしさを推定するということが可能である。しかし、これでは、連続発声において女性らしい発声ができているかは判定できない。また、母音ごとに分析を行うということは、入力としてどのような内容でも受け付けるわけではないため、システムとして実装した場合の利便性も低くなる。そこで、本論文ではテキスト非依存の話者照合技術を応用することで、この問題の解決を図る。

また、線形回帰による予測という手法を用いることで、コンピュータによって得られた音響的女声度から、聴取実験によって得られた声の女性らしさ（知覚的女声度）の推定を行う。また、この技術をインターフェースとして実装し、簡易に利用できるようにし、ボイスセラピーの中で使用していただき、実用性を検討する。

1.2 本論文の構成

まず、第2章で、性同一性障害とそれにまつわるボイスセラピーや声の問題について詳しく記述する。次に、第3章では本論文と関連する音声工学の基礎について、第4章では、音声に含まれる非言語的な情報についての概観と、声の性別、すなわち声の男性らしさ・女性らしさについての先行研究をまとめる。第5章では、話者認識技術とそれを応用した先行研究を紹介する。

続いて、第6章で、聴取実験を行い、性同一性障害の音声が、人間によってどの程度女性らしいと知覚されるか調べ、知覚的女声度を定義する。第7章では、話者照合技術に基づいて、コンピュータによる音響的女声度を算出する方法を提案する。第8では、実験の結果を述べ、この2つの女声度（知覚的女声度、音響的女声度）の相関などから提案手法の精度や妥当性を検討する。第9章では、第6～8章で提案した手法をもとに、知覚的女声度推定器としてインターフェースの実装を行い、実際のセラピーにおける臨床応用の様子を報告する。

最後に第10章で結論を述べる。

第2章 性同一性障害

2.1 性同一性障害

性同一性障害 (Gender Identity Disorder, GID) とは、医学的な疾患のことで、もって生まれた性別、すなわち生物学的な性別と、自分が認識する性別が異なる症状のことを指す [1]。生物学的には男性でありながら、自分は女性であると認識し、女性になりたいと望むといったことが代表例である。以下、性同一性障害についての詳細を述べる。

2.1.1 性同一性障害の定義

性同一性障害は米国精神医学会が定めた診断基準 DSM-IV-TR (Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision) [2] によると、以下のような 4 つの診断基準を満たす場合に診断される。

1. 反対の性に対する強く持続的な同一感。
2. 自分の性に対する持続的な不快感、またはその性の役割についての不適切感。
3. その障害は、身体的に半陰陽を伴ったものではない。
4. その障害は、臨床的に著しい苦痛または、社会的、職業的または他の重要な領域における機能の障害を引き起こしている。

まず、1 つ目の基準として、自分の生物学的な性とは反対の性に対する、強く持続的な同一感がある。たとえば、MtF (Male to Female、男性から女性へ性別を移行するもの) であれば、反対の性である女性としての考え方や行動をとることや、手術などをしてでも女性になりたいと考えたり、女性としての社会的生活を送りたいと感じるということが挙げられる。MtF とは反対に、生物学的には女性でありながら男性になりたいと希望するものは、FtM (Female to Male) と呼ばれている。

2 つ目の基準として、自分の性に対する持続的な不快感や、その性の役割に対する不適切感がある。MtF であれば、声が低いことや、体格、男性生殖器を所持することへの不快感や、あるいはスーツ姿になることに抵抗を感じるといったことが挙げられる。

つまり、性同一性障害とは「自分が男であるか女であるか」という、性別に関する自認、すなわち性自認 (gender identity) の問題であるということが出来る。

2.1.2 同性愛などとの混同について

混同されやすい問題として、同性愛などの性的指向 (sexual orientation) の問題や、半陰陽 (性染色体、性腺、内性器、外性器などの身体的な性別が、非典型的な状態) の問題などが挙げられるが、これらと性同一性障害は本質的に異なる問題であり、区別されている。

たとえば同性愛者についてみた場合、男性同性愛者では肉体的にも性自認においても男性である。MtF のように、女性になりたいという意味は持っているわけではなく、男性として男性に性的指向を持っているということになる。性同一性障害者の場合では、自認する性と反対の性に対し性的指向を持つことが多い。つまり、MtF であれば、男性に対して性的魅力を感じる人が多い。ただし、MtF であっても女性やあるいは両性に対して性的指向を持つこともある。

2.1.3 日本における性同一性障害

日本では、1995 年に埼玉医科大学の倫理委員会において、はじめて性同一性障害の正式な治療が認められた。それまでは、性同一性障害に関して正式な治療は行われておらず、自己の性別への違和感に苦しむ人達は、性同一性障害に対する差別と偏見に苦悩する日々を送ってきた。日本における性同一性障害の正式な患者数はわかっていないものの、日本では埼玉医大が診療を開始した 1996 年から 2003 年の間に、約 2200 名の患者が各地のジェンダークリニックで診察を受けている。診察を受けた患者が全て性同一性障害者であれば、MtF と FtM をあわせて 6 万人に 1 人ほどの患者が存在するということになる。なお、諸外国ではおおよそ MtF が 3 万人に 1 人、FtM が 10 万人に 1 人という統計があるが、これは国によってさまざまである。

また、2003 年 7 月には「性同一障害者の性別の取り扱いの特例に関する法律」(いわゆる特例法) が成立し、2 名以上の精神科医に「性同一性障害」と診断され、以下の条件に該当する場合に限り、社会的性別の取り扱いの変更が認められるようになった。

1. 20 歳以上であること
2. 現に婚姻をしていないこと
3. 現に子がいらないこと
4. 生殖腺がないこと又は生殖腺の機能を永続的に欠く状態にあること
5. その身体について他の性別に係わる身体の性器に係わる部分に近似する外観を備えていること

このように、日本における性同一性障害者を取り巻く環境は、ここ数年で劇的に変化している。2004 年に特例法が施行されてからは、自己の生物学的性別をカムアウト (周囲に告知) することなく、希望の性別で日常生活が送れるようになっており、性同一性障害者の QOL (Quality of Life) は飛躍的に向上したといえる。

テレビドラマにおいても性同一性障害をかかえる生徒が中心人物として登場した例や、性同一性障害者が地方議員選挙に立候補し当選したときにも、社会的な非難はほとんどみ

られなかったことなどから、性同一性障害にたいする偏見は薄らいできているとみることができる。

また、性別違和を訴える子供に対する親や教育者の意識も変化している。2005年4月、神戸で6歳時に性同一性障害と診断された少年が、女兒として小学校に入学したというケースがある。この少年は、思春期に再判定を受ける予定であるものの、現在のところ女兒として名簿に載っており、また、女子トイレを使用している。

2.2 MtF と声の問題

性同一性症害者が抱える問題は多く存在するが、以下の二つが問題となることが多い。

1. 外見・容姿の問題
2. 声の問題

容姿の問題については、ホルモンによる薬物治療や美容整形による外科的治療によって、ある程度の変化が望むことができる。これは、MtF と FtM で大きく変わらない。一方で、声に関する問題に関しては事情が異なってくる。女性から男性へ性別の移行を希望する FtM の場合、男性ホルモン投与で声の男性化はおおよそ達成される [3][4]。しかし、MtF に関しては、ホルモンや声帯手術による声の女性化の効果はほとんど得られない [5][6][7]。また、MtF 当事者からの訴えには以下のようなものがある。

1. 外見では女性とされていたのに、声を出した途端、生物学的な性が男性であると認知された。
2. 相手が、外見からでは自分が男性か女性かわからないと思われた時は、挨拶など声かけなどにより、こちらの声を聞こうとすることがある。

このような例から、我々が性別を判断を行う際には声に頼る部分が大いといえることができる。

この声の問題は、外見が完全にパス（周囲に女性とされている）している MtF ほど深刻である。外見がパスしていなければ、声は元の性別を確認する手段にすぎない。その反面、外見からは完全に女性と認識されている場合には、女性化されていない声を出した後にはその声を聞いた相手から、女性であると思っていたが実際には男性であった、という発言をされることがあり、傷ついたという訴えも少なくない。また、自分の声が他人にどう聞こえているかわからず、不安に思うことや、また声に自信がないために他人と話すことを躊躇し、内に引きこもってしまう場合もある。MtF 当事者が、希望の性でうまく社会適応して生きるためには、音声によるコミュニケーションが存在し、それがコミュニケーション手段の中心であり続ける限り、声の問題を避けて通ることはできない。このように、声の問題は MtF 当事者達の QOL を下げる大きな要因となっている。

2.3 声帯手術による声の女性化の試みとその問題点

このように、MtF 当事者によって声の問題というのは、自己が希望する性別で生活し、社会に適応していくためには必要不可欠であるといえるわけであるが、MtF の場合は、FtM に対して有効であるようなホルモン治療は効果が期待できないことがわかっている。声の問題を解決する方法としては、そのほかに声帯手術が挙げられるが、これはどの程度の効果がえられるのであろうか。

櫻庭ら [8][9] は、第三者に、話者に MtF が含まれていることを知らせずに音声を聞かせて、話者の性別を男女の 2 択で判断させる聴取実験を行っている。この聴取実験では 1 話者につき、25～45 名の聴者が話者の性別や年齢を推定している。櫻庭らが集めた 4 名の声帯手術施行者の音声に対する聴取実験結果を見てみると、女声と判定された割合は 0～50 % (平均 15%) にしかなかった。

このように声帯手術を行ったとしても、女声を獲得することは困難である。この理由としては、現在行われている声帯手術が、声帯の厚みを薄くすることにより、声の高さ (基本周波数) を高くすることのみを目指していることが考えられる。この術式では、声道形状の変化はまったく望めない。第 3 章でも説明するが、声の生成に関連する項目は、基本周波数に関連する声帯波形だけではなく、その声帯波形に変化を加える声道形状も存在する。声道の長さや大きさは男女で異なるため、この部分に対する考慮を行わなければならない声帯手術だけでは声の女性化は望めないといえる。

2.4 トランスセクシュアルボイスセラピー

櫻庭は男性から女性へ性別の移行を希望する性同一性障害者 (Male to Female transgender / transsexual = MtF) に対して、声を女性化させるためのトランスセクシュアルボイスセラピー (Transsexual Voice Therapy) を行っている。櫻庭らが行っているボイスセラピーは、

1. 声の評価
2. 発声訓練
3. 精神的ケア・カウンセリング

から構成されており、個人に応じた個別プログラムを組んで実施されている。

2.4.1 声の女性らしさ決める要因

MtF が女声を獲得するためには

1. 高過ぎず低過ぎない声の高さ
2. 喉を絞ることによる声道形状の適切な制御
3. 女性らしい話し方

が必要であり、ボイスセラピーでもこれらを獲得できるような指導が行われている。

2.4.2 声の高さ

女声の獲得のために、声の高さが重要であることはさまざまな研究から知られている[10][11]。ただ、裏声やいわゆるアニメ声などのように、単に生物学的な男性が男声のまま基本周波数を上げて発声しただけでは、女声と判定されるようにはならない。櫻庭らの聴取実験[8][9]では、女声判定率は30%ほどしかなく、MtF本人の満足度に比べて、第三者が女声と認識する率は低いことがわかっている。また、この聴取実験では、180～230Hzあたりの声をもっとも女性らしく聞こえる高さであることが示された。

2.4.3 声道形状の制御

そのほか、声の高さをあげる場合、男声をそのまま裏声にすると不自然な声に聞こえてしまう。ボイスセラピーでは、喉を絞るようにして声道の形状を少し変形させて、地声の一番高いところを引き伸ばすようにすると、自然な声に聞こえやすいといった経験上の知見が得られている。

第2.3節で示したように、声帯手術によって声の高さを上げただけでは聴取実験によって女性と判断される率は50%以下であった。それに比べて、80%以上を超えて女声と判定された音声を持つMtFは、もともとの声質が女声に近いものや、狭義・広義のボイスセラピーの経験者であった。ボイスセラピーでは、喉を絞るようにして声道形状を変化させた上で、高い声を出す訓練を行う。そのため自然と喉仏の位置が上がり、声道形状の変化を起こすと考えられる。この発声法によってMtFの声道形状は生物学的な女性の声道に近い形になり、聴取実験において声が女性であると判定される率が上昇すると考えられる。

2.4.4 女性らしい話し方

女声獲得のために、声の高さや声道形状の制御と同様に重要なものが、女性らしい話し方である。女性らしい話し方の特徴として

1. 語尾を延ばす
2. 語尾を下げない
3. 抑揚に富む
4. 軟起声を使用する
5. 鼻音化する

などが挙げられる。しかしながら、極度に技巧的な話し方をした場合だと、男性がわざわざ話しているようにしか聞こえてしまう。このような場合だと、聴取実験によって女声と判定される割合は30%程度となってしまう。

むしろ、女性らしく話す技巧はなくても、フルタイムで女性として生活し、職場や地域で女性グループの中に溶け込んで生活をしている人の方が、女声と判定される割合は高くなる傾向にある。話し方というのは、性差による文化のようなものであり、女性文化に浸ることによって始めて獲得が可能になるものと考えられる。

2.4.5 女声を判断する評価者の問題

MtF の場合、声の女性化のためにボイスセラピーに基づく訓練が必要となるが、性転換者特有の難点が存在するのも事実である。

MtF 当事者、あるいは MtF 当事者に関わっている専門家の間にしばしば生じる現象として、自分の持っている声を女性であると判定する基準値がずれてくることがある。MtF の元の性別を知っていると、声が女性であると判定する基準が必要以上に厳しくなったり、反対に甘くなったりしてしまうことがみられる。また、多くの当事者に接しているセラピストには、そのことによる偏見や慣れが生じてしまい、かえって声の女性らしさについての正確な判断がしにくくなることもある。また、セラピストによって女声の判断基準が違ってくる場合も生まれてくる。ある病院でボイスセラピーを受け、セラピストからは完全に声が女性化したと判断された MtF の声であっても、第三者の聴取判定にかける場合には、依然として男性であると判断されることもあった。同様の問題は音声合成の開発においても知られた事実である。特定の合成音声を繰り返し聴取することで、その主観的品質が自ずと向上してしまう、という問題である。知覚の馴化は不可避免的な現象であるため、生身の人間が判断する以上、その解決は難しい。

そこで、声帯の外科的手術の効果、あるいはセラピーによる進捗度や成功度を測るためには、私情や偏見を持たずに、当事者の声を判断してもらう第三者による聴取実験は必須だとなってくる。しかしながら、実験の意図を知らず偏見を持たない第三者に判定してもらうためには、絶えず新しい聴者を求める必要が生じる。また、知覚実験における判断には個人差がみられるため、統計にたえうるだけの聴取者の人数を確保もしなければならない。このため、第三者による聴取実験によって男女声の判定が行われるまでには、音声収録が行われてから相当の時間が経過してしまうことが多い。そのため、簡易かつ先入観なく聴取実験のシミュレーションを行えるシステムが望まれている。このような聴取実験や評価作業の難点を解決する方策の一つとして、私情や偏見をはさまずに評価可能な、コンピュータによる知覚的な性別の判定が行えるようになることは、臨床的意義があると考えられる。

第3章 音声生成の原理と工学的モデル

3.1 ソースフィルタモデル

現在の音声工学において、音声生成の過程は、図 3.1 に示されるようなソースフィルタモデルで近似される。声帯で生まれた音源波形が、声道のフィルタを通り変化し口唇から放射されたものが観測される音声波形となる [12]。

音源波形は一般に非対称の三角波で近似される。最も重要な情報はその基本周波数である。波形の形状は、スペクトルの傾斜として表れ、 -6db/octave の周波数特性で近似されることが多い。一方、声道のフィルタは、全極型の共鳴管として近似される。音の放射特性に関する考慮も必要である。これは $+3\text{db/octave}$ の周波数特性で近似されることが多い。そのため、放射特性と声帯波形の二つの周波数特性をキャンセルし声道のスペクトルを得るために、 $+3\text{db/octave}$ の高域強調を行うことが多い。

3.2 音響特徴量

音声から得られる音響的な特徴量は大きく分節的特徴と韻律的特徴に分かれる。これらの違いは、前述の音声生成のメカニズムから導かれるものである。以下でそれぞれについて詳しく述べる。

3.2.1 分節的特徴

分節的特徴は、音素など個々の音を特徴付ける特徴である。これは、声道形状の特徴、すなわち全極型の声道フィルタとして近似されるスペクトル包絡の情報に相当する。分節的特徴として用いられるものとして、具体的には、フォルマント、ケプストラム、MFCC (Mel Frequency Cepstrum Coefficient) などが挙げられる。

フォルマント

フォルマントはスペクトル包絡において極大となっている山状の部分のことで、その中央周波数と帯域幅で定義することが多い。また、一般にスペクトル包絡は複数のフォルマントから構成され、周波数の低いほうから順に第 1 フォルマント (F_1)、第 2 フォルマント (F_2) と呼ばれる。フォルマントは声道の共鳴周波数を表しており、声道の長さや太さ、下の位置や口の形などの調音によって変化する。一般に声道の長さに反比例して低くなるといわれている。

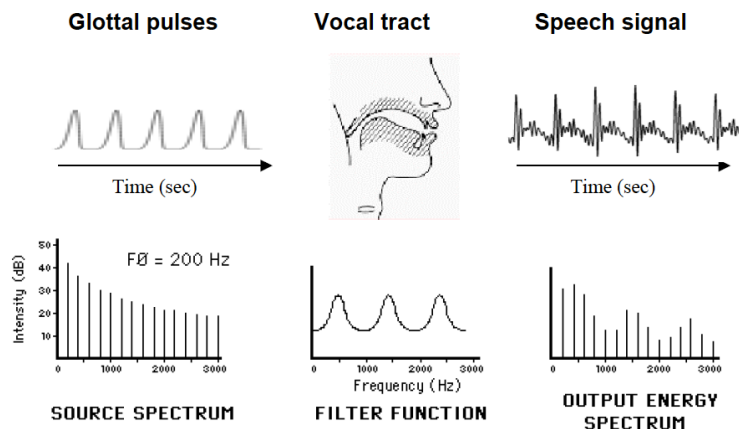


図 3.1: ソースフィルタ音声生成モデル

低次のフォルマントは、調音の影響を大きく受け変化し、音韻による音の聞こえの相違を生み出す。一方で、フォルマントの高次のものは、音韻性による影響はあまりみられず、個人によって変わってくるといわれる。そのほか、スペクトル包絡の長時間平均をとれば、音韻性が平均化され個人性などを表すようになる。この特徴は話者の認識に有効であると考えられる。

フォルマント周波数は線形予測分析により求めることができる。

ケプストラム

近年ではスペクトル包絡のをあらわす特徴としてはケプストラムがよく用いられる。これは対数パワースペクトルを逆フーリエ変換したもので、低域にスペクトル包絡、高域に音源情報が近似的に分離する。低域の部分のみを用いることでスペクトル包絡の情報を得ることができる。

3.2.2 韻律的特徴

韻律的特徴は音声全体の抑揚に関するもので、音源で生み出される情報を表す。実際に用いられる特徴量としては、基本周波数 (F_0) や声の大きさ (パワー) や、それらの時間的パターンが挙げられる。たとえば基本周波数のパターンは、図 3.2 のようになる。

一般に、韻律的特徴は、発話の内容との関係は薄く、音声認識で用いられることは少ない。日本語の場合では、単語レベルでは声の高さによるアクセント型が存在し、単語を区別するが、現在主流の音声認識では、声の高さは用いられていない。ただし、中国語のような声調言語においては、音の高さやその変化が非常に重要であり、声の高さをを用いた音声認識も行われている。

一方で、パラ・非言語情報、特に感情などは韻律的特徴との関係が強いとされ、コンピュータによって情報処理をする場合は、韻律的特徴を見ることで、感情などの推定を行っていることが多い。しかし、実際に分析や認識に用いる場合、基本周波数やパワーの時間

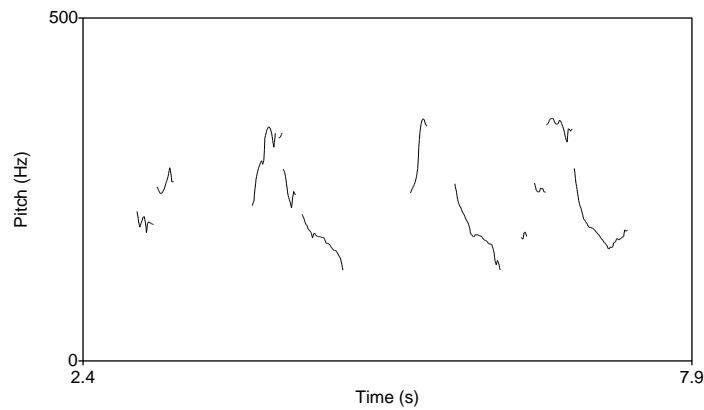


図 3.2: f0 パターンの例

的パターンそのものの情報を扱うことは難しい。そのため、発話全体から、さまざまな統計量を計算し、それらをあわせて用いることで認識を行っている。基本周波数やパワーのに関連する統計量としては、それぞれの平均値、中央値、分散、最大値、最小値、レンジ、第一四分位、第三四分位、ジッタ、回帰直線の傾きなどが挙げられる。あるいはこれらを最大値や平均などで正規化したものも用いられることがある [13][14][15][16]。

第4章 音声に含まれる情報と声の女声らしさと先行研究

4.1 音声に含まれる情報

音声にはさまざまな情報が含まれているが、それらは大きく言語情報とパラ・非言語情報の2つに分けることができる。言語情報は、文字で表現できる情報に相当する。一方、パラ・非言語情報は文字で表現できないような情報に相当する。これは、たとえば、性別、年齢、感情といったような情報である。パラ言語情報はその中でも話者が制御可能なものであり、感情が代表的なものである。それとは逆に、非言語情報は話者が制御不可能なもので、性別、年齢などの情報がそれに相当する。これをまとめたものが表 4.1 である。

これらの情報を対象とした音声情報処理も行われている。たとえば、音声に含まれる感情を認識しようとしたものが音声感情認識であり、話者性を取り出して話者を識別しようというものが話者認識・話者照合である。

4.2 声の女性らしさ・性別に関する先行研究

声の女声らしさは、前節で述べたうち非言語情報に属する情報である。声の性別に関する研究は古くから行われているが、ここでは過去 40 年程度の研究について取り上げ説明する。

4.2.1 声道模型を用いた合成音声による研究

梅田ら (1966)[17] は音響的声道模型を用いた音声合成をもとにして、日本語 5 母音に対応する声道模型 (図 4.1) の基準化を行った。すなわち、母音ごとに標準的な声道の形を求めたものであるが、同時に声道形状の男女差や年齢差を考え、複数の声道長の模型を用いた。本稿ではこの部分を取り上げる。

声道模型は、男のモデルとしては、声道長 17.5cm のもの、女のモデルとして声道長 14cm のもの、子供のモデルとして声道長 9cm のものが使われた。各声道模型に声帯波形を入力

表 4.1: 音声に含まれる情報

言語情報	文字で表せる	-
パラ言語情報	文字で表せない	制御不可能
非言語情報	文字で表せない	制御可能

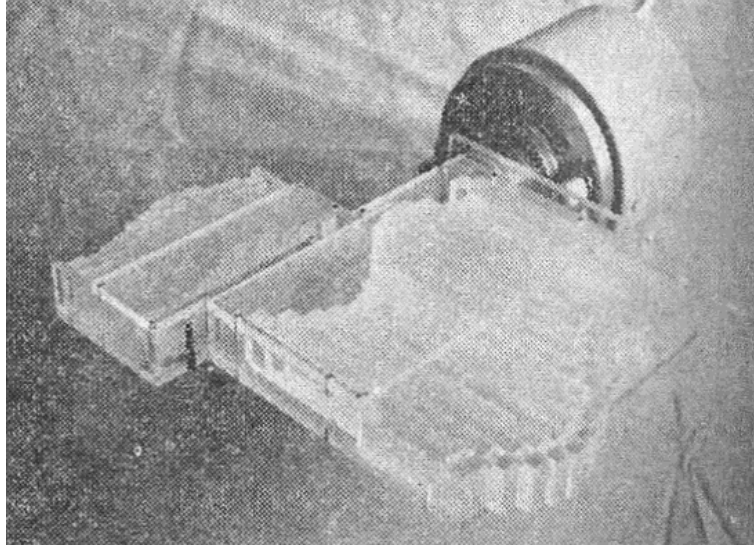


図 4.1: 声道模型

することで音声合成されるが、自然な音声にするためには、模型の声道長にあわせて、適切な基本周波数の声帯波形を入力する必要がある。

声道長と基本周波数の組み合わせにより、男性、女性、少年、子供の声として聞こえる範囲が示されている（図 4.2）。示されている領域のいずれにも含まれない場合、人間の声に聞こえなくなり、場合によって、声の明瞭度も失われる。

また、基本周波数だけでなく、入力する声帯波形の形に声質が影響を受けることも示されている。声帯波形は鋸状の波形（非対称三角波）で表されるが（図 4.3）、その立下りによってスペクトル包絡が決まる。男女や子供で立下りの時間は異なるため、音声のスペクトル包絡にもそれぞれの特徴が現れる。立下りが急峻になるにつれ張りのある声になり、女性らしさが失われていき、その場合、声が高い場合には子供のように、声が低い場合には少年のように聞こえることが示されている。

4.2.2 スペクトルの“山”を用いた声の男女の研究

佐藤 (1974)[18] は日本語五母音について母音別に女性らしさの分析をした。日本語の書く母音について、スペクトルの“山”で特徴づけ、その山の中心周波数と帯域幅がどの範囲にあるときに聴取実験において女声と判定されるかが調べられた。ここでいうスペクトルの“山”は、基本的には2つのホルマントから定義されるものであり、/u/を除く4つの母音は2つの山によって特徴づけられる（図 4.4）。/u/に関しては、フォルマントが等間隔に現れるため、山は定義していない。以下に紹介する実験では、合成された音声を聴取して女性に聞こえるかどうかを判断し、パラメータを決定しているが、詳細な実験条件は示されていない。

そのほか、スペクトル包絡の傾斜や基本周波数が声の女性らしさに与える影響も示されている。周波数に関しては、女性と判定される範囲は 265 ~ 300Hz であった。ここで、基

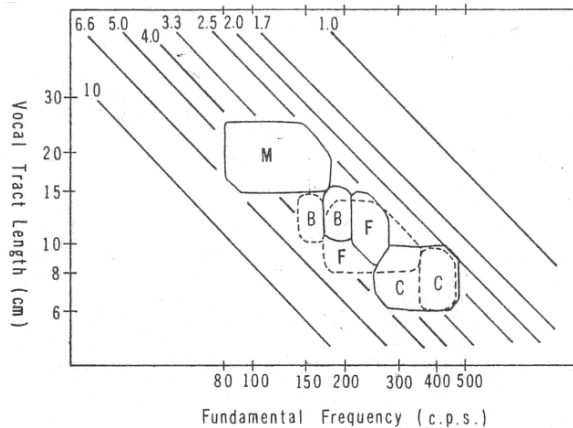


図 4.2: 声道長と基本周波数のマッチング

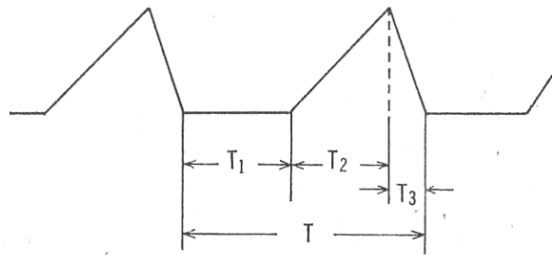


図 4.3: 音源波形

本周波数の影響を見るときには、スペクトル包絡を決めるパラメータについては女性に聞こえる値に固定している。ただし、適切な基本周波数は母音によって異なる。

スペクトル包絡の傾斜が女性らしさそのものに与える影響は少なかった。また、スペクトルの山の位置や帯域幅の適切な変換の仕方は母音ごとに異なっている。このため、男女の声の違いは単純に声道長だけでは説明できないことが分かった。

4.2.3 男声連続音声の女声への変換する実験

安広ら (1976)[19] は線形予測分析合成を用いて、入力した男声を女声に変換する実験を行った。

男声の入力に対し、母音の種類に関わらず、ホルマントの中心周波数と帯域幅、基本周波数を一定の割合で変換した。ホルマントに関しては 1.3 倍、基本周波数は 2.1 倍とした。

それに加え、声帯波形の差を補正するため、コサイン櫛状のフィルタを通す処理をが施された。これは、声帯で発声する三角波において、立上がりと立下りの急峻さが男女で異なることにより生まれるスペクトル包絡形状の差を補うものである。

以上により男声の連続音声を女声に変換したものを聴取した結果、出力の音声は女声として聞こえるが、やや音韻性に変化がみられた。ただし、音韻を聞き間違えるほどの変化

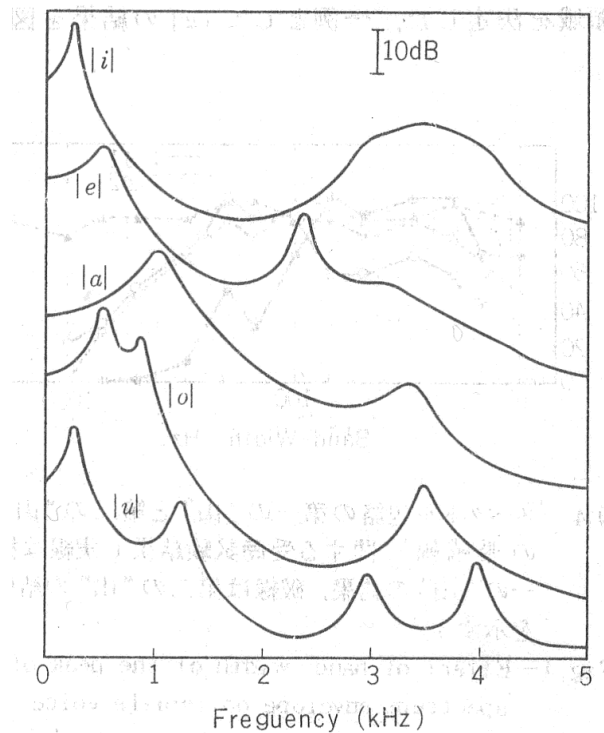


図 4.4: 各母音のスペクトル形状

ではなかった。

4.2.4 声道面積関数をパラメータとし主成分分析を用いる手法による研究

太田 (1977)[20] は日本語定常五母音について、声道面積関数をパラメータとして用い、主成分分析により韻質（音韻性）と声質にがどのように表されるかについて分析した。各母音の声道断面積は図 4.5 のように表される。この図で左の列が男性、右の列が女性の例である。

これによれば、日本語五母音は、韻質 3 次元に加え（男女）声道長差 1 次元分をあわせて計 4 次元で表される。男女声道長差は、/a/, /u/, /o/ については第 1 軸、/i/, /e/ については第 2 軸として現れており（図 4.6）、声道長の差に基づく男女声の違いを完全に 1 つの軸に集約できたわけではないようである。なお小児の音声を混在させた場合には、小児の音声は女性よりさらに声道が短い方向の領域に現れた。

4.2.5 声の性別の知覚に関する先行研究

John W. Mullennix ら (1995)[21] は、人間が声の性別をどのように知覚しているかについていくつかの実験を行った。

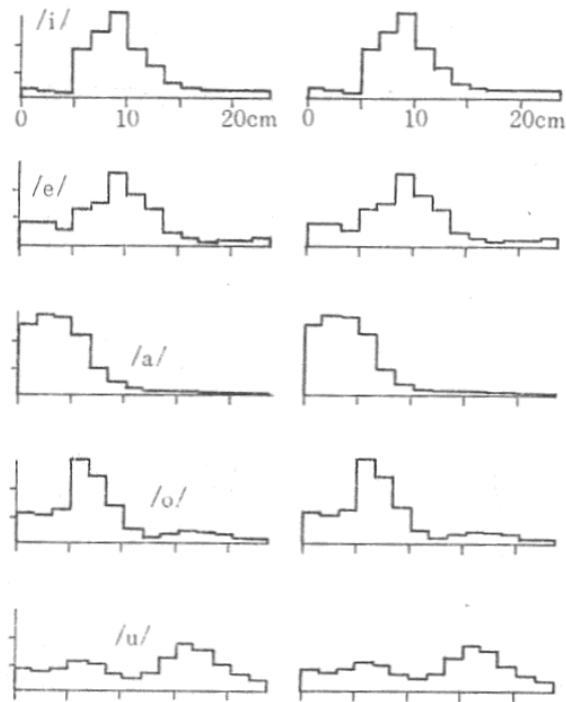


図 4.5: 各母音の声道断面積関数

声の性別に関する知覚がカテゴリー的かどうか

ひとつ目の実験は、人間の声の性別に関する知覚がカテゴリー的かどうかを調べるものである。刺激音声は音声合成によって作られた母音/i/で、男性的な声から女性的な声まで知覚的に広がりを持つ 11 の音声からなる。予備実験によりさまざまな F0、フォルマント周波数、声帯波形の組み合わせを持つ音声を聴取してもらっており、最も女性らしい音声と、最も男性らしい音声を取り出して端点とし、F0 とフォルマント周波数を線形に補間することで中間の音声を合成した。声帯波形については、呼び実験では知覚に与える影響はほとんど見られなかったため、すべての合成音声について同一のものが用いられた。

これらの音声に対して最も男らしいものを 1、最も女らしいものを 11 として、番号が 2 つ違う音声のペア（たとえば 2 番と 4 番の音声ペア）に対し、ABX 法で識別をする実験が行われた。また、各音声に対し 6 段階で女性らしさ男性らしさを評点させた。聴取者の人数は 30 人であった。

この実験の結果は図 4.7 のようになっており、刺激音声の性別の連続的な変化に対して、識別率のピークや急激な評点の変化はみられず、人間の合成音声の性別に対する知覚はカテゴリー的ではないことが示された。

第三のカテゴリーが存在するかどうか

第 4.2.5 節の実験では、音声の性別に関するカテゴリー的な知覚は見られなかったが、中間的な第 3 のカテゴリーが存在する可能性がある。

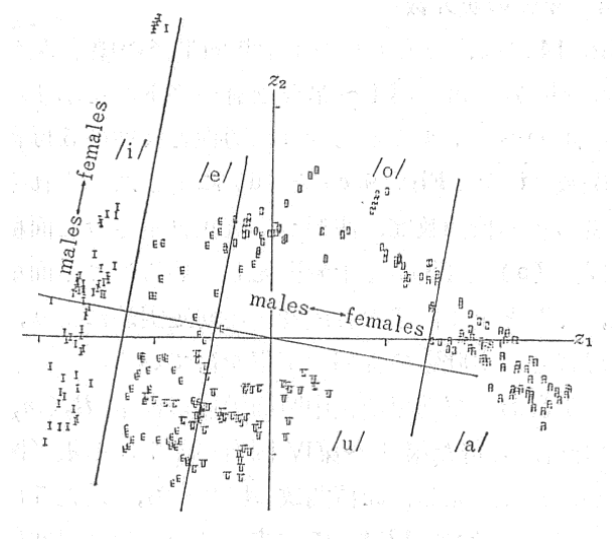


図 4.6: 声道断面積関数の主成分分析の第 1 軸と第 2 軸に対する各母音と男女の広がり

そこで、第 4.2.5 節と同様にして聴取実験が行われた。ただし、今回は 6 段階の評点ではなく、男、女、その他の 3 つのカテゴリのどこに聞いた音声属するかを回答した。

その結果は図 4.8 のようであった。中間的な音声ではその他のカテゴリに属する傾向があるが、その他のカテゴリの存在を裏付けるほどではなかった。

その他

そのほか、同時に男女判定の基準となる音声を聞かせて評点をつけさせる実験も行われた。ここでは基準として提示される音声によって評点が有意に影響を受けることが示された。

4.2.6 性別認識に関する先行研究

Ke Wu ら (1991)[22] は、音素の種類ごとに使用する音響パラメータなどを変えて、音声から性別を認識する実験を行った。

使用する音声は、母音（二重母音を除く）、無声摩擦音、優勢摩擦音であり、使用するパラメータは、LPC、自己相関係数、ケプストラム、反射係数、フォルマントと基本周波数のセット（ただし母音を用いる場合のみ）であった。

男女のテンプレートは基本的に平均をとることで作成された（図 4.9）。テキスト非依存の手法であり、時間情報は用いられていない。たとえば母音の LPC のテンプレートであれば、発話ごとに LPC の 6 フレームの平均をとることである母音のテンプレートを作り、つぎにすべての母音の平均をとることである話者の母音のテンプレートとなり、それを全話者の平均をとり、性別ごとの母音のテンプレートとなる。

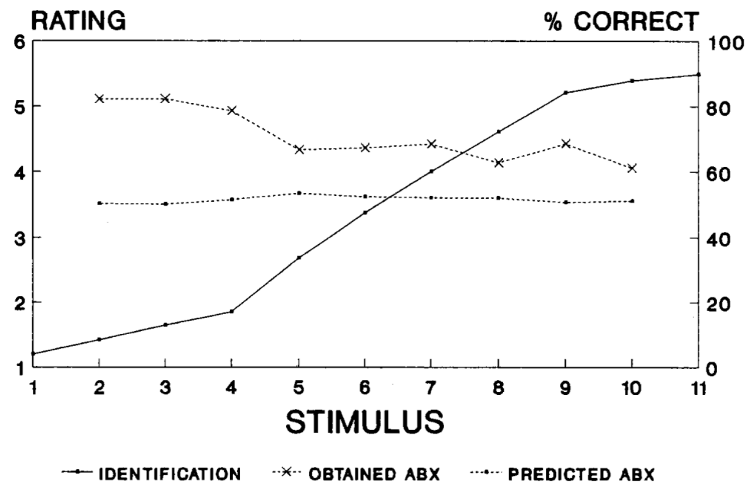


図 4.7: 声の性別の連続的变化に対する知覚実験結果

またテンプレートと入力音声の距離尺度も複数試されている。ユークリッド距離、ガウス分布の確率密度関数のほか、LPC では板倉距離、ケプストラムではケプストラム距離も実験している。

全体的に正解率が高かった組み合わせは、パラメータとして反射係数またはケプストラムを用い、距離尺度としてユークリッド距離を用いた場合であった。特に、反射係数とユークリッド距離の組み合わせでは、母音の性別認識においては 100% の精度を得た。

4.3 声の女性らしさに関する先行研究のまとめ

上に挙げた研究例をみると分かるとおり、既存の研究では音声としては母音などの孤立発声を対象としたものがほとんどである。それらは母音であれば、各母音ごとの分析を行っている。パラメータとして用いる音響特徴量も、基本周波数やフォルマント周波数を用いるものが主流であり、ケプストラムなどの特徴が用いられるようになったのは、近年になってからである。これらのことをふまえると、とくに実際に使用するような応用を考えた場合に、連続発声に対して、発話内容を問わず、また安定した分析に基づいた新しい枠組みが必要になる。

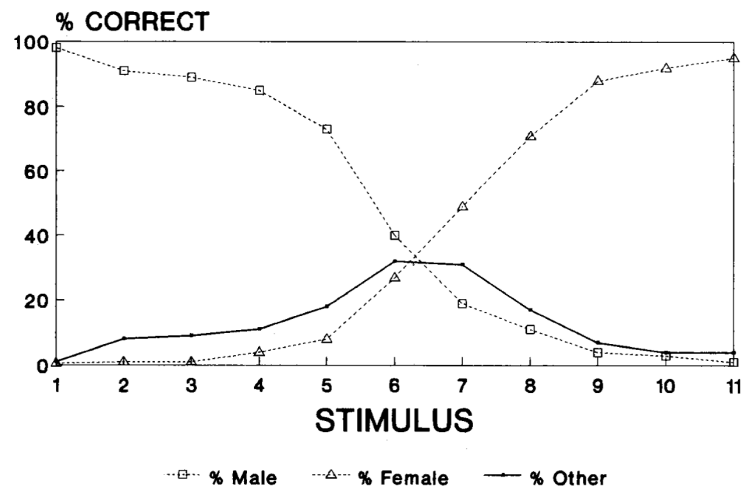


図 4.8: 3 カテゴリーでの知覚実験結果

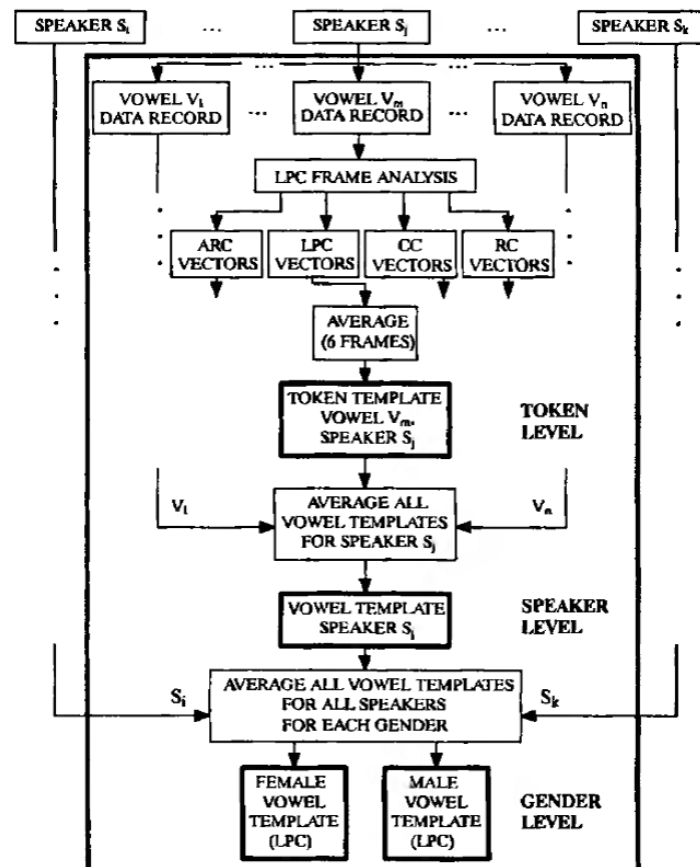


図 4.9: テンプレート作成手順

第5章 話者認識技術

5.1 話者認識

本論文では、コンピュータによる声の女性らしさの推定は、話者認識の技術をベースとしている。話者認識には、大きく分けて話者識別と話者照合の技術が存在し、また発話内容を利用するものや発話内容に依存しないものなどがある。ここでは話者認識についての概要を説明する。

5.2 話者認識の種類

話者認識とは、音声から個人性を抽出しその情報を基にして、話者の特定などを行うことである。話者認識とよばれているものには、話者識別と話者照合という2つのものがある[24]。

話者識別では、ユーザが発した音声を入力し、その音声を分析し、誰が発声したものであるかを判定する。あらかじめ登録されている人物の中から一番適合する人物を正解として出力することになる(図5.1)。

それに対し話者照合は、ユーザが音声を発すると同時に、自分がだれであるかを入力同時に入力する。システムは入力音声、ユーザの主張する話者が発したものであるかを判定する(図5.2)。まず話者識別を行った後、話者照合により本当に話者識別により示された人物であるかを照合する、ということも行われる。

5.3 テキスト依存性

また、認識の技術としては、テキスト依存(text-independent)のものと、テキスト非依存(text-dependent)のものに大きく分けられる。テキスト非依存のものは、発話内容を制限しない方式で、どのような内容であっても入力を受け付ける。テキスト依存のものは、発話内容があらかじめ決まっており、内容に特化した処理系を構築する。

話者照合という特性を考えれば、実用性はテキスト依存な仕組みにすることが可能である。すなわち、システムがユーザーにテキストを表示し、ユーザーはそれを読み上げるという形にすることは難しくない。一方、テキスト非依存の方は、認証の精度を考えた場合、問題はより困難となるものの、より柔軟で安定したシステムを作ることができる。

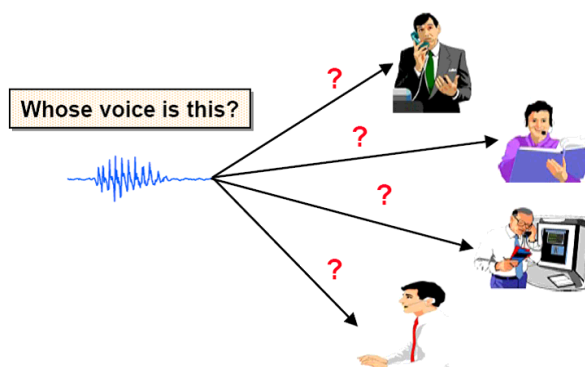


図 5.1: 話者識別

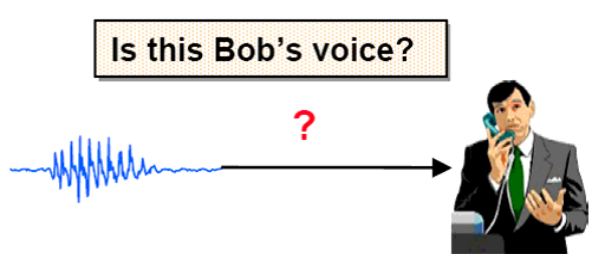


図 5.2: 話者照合

5.4 認識手法

実際の話者認識のシステムはさまざまであるが、ここでは GMM (Gaussian Mixture Model) ベースの話者照合技術を主に説明する [25]。

5.4.1 GMM ベースの話者照合

ある話者を照合するシステムを作る場合には、その照合の対象となる話者の音声から作られたモデルと、その他のすべての話者の音声から作られたモデルを用意する。音声データは発話全体を用いて分布化されるが、これにより音韻性の違いなどは吸収され、音声における静的な要因が強く現れることになる。

話者 s のモデルを M_s とした場合、話者照合 (観測された音声 o が話者 s の音声であるか否かを判定する) では、

$$\Lambda = \log P(o|M_s) - \log P(o|M_{\bar{s}}) \quad (5.1)$$

を算出する [26][27] すなわち、入力した音声の特徴量の系列に変換した後に、各モデルから生成される確率を計算し、その対数尤度の差を求める。その対数尤度の差 Λ が、あらかじめ決められた閾値よりも大きければ、音声 o は話者 s の音声であると判定 (受諾) し、小さければ拒否する。これを図 5.3 に示す。使用する特徴量はケプストラムなどスペクト

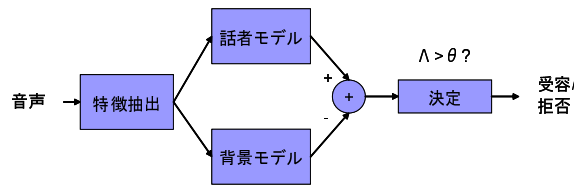


図 5.3: GMM ベース話者照合の概略図

ル包絡に関するものが多い。また、モデルを作る際には、学習データの音声に含まれる無音区間や無声子音区間は、話者性と関連が薄いため取り除かれる。スペクトル包絡をベースとしたモデルの場合、話者の特徴が現れるのは有声音の部分であるからである。

5.4.2 その他の手法を用いた話者認識

近年では SVM を用いた認識も行われるようになってきている [28]。テキスト依存の話者照合の場合、DTW(Dynamic Time Warping) により時間軸方向に非線形圧縮・伸長をし、マッチングを行うという手法もある。また、韻律的特徴を用いた話者認識も検討されている。

5.5 性能の評価

話者照合の精度は、False Reject - False Alarm 曲線で表されることが多い。入力した音声を発した人物が、主張する人物と同一であるかどうかを判定するときの閾値を変えることで、同一であると判定すべきところを拒否してしまう場合 (False Reject) と、同一でないとすべきところを許容してしまう場合 (False Alarm) の確率が連動する。この曲線を見ることでシステムの優劣を知ることができる。

5.6 話者照合を応用した先行研究

話者照合を応用した研究として、年齢の自動推定に関する研究を取り上げる。

5.6.1 年齢推定

関口ら [29] は従来の音声認識の技術を応用して、音声から話者の年齢を推定する試みを行った。ただし、この研究で推定しているのは実年齢ではなく知覚的な年齢である。これは、年齢情報を伴った音声データベースが少ないということと、また、人間の判定する年齢は常に知覚的なものであるからという理由からである。

用いた音声データは子供音声、JNAS(日本語新聞読み上げ音声)、SJNAS(Senior JNAS)である。これに対し 30 名の聴者に何歳に感じるかというラベルをつけてもらいそのデータを下に年齢の推定を行った。ただし、用いる音声データの都合上、知覚的な年齢の存在する範囲に偏りがある (図 5.4)。

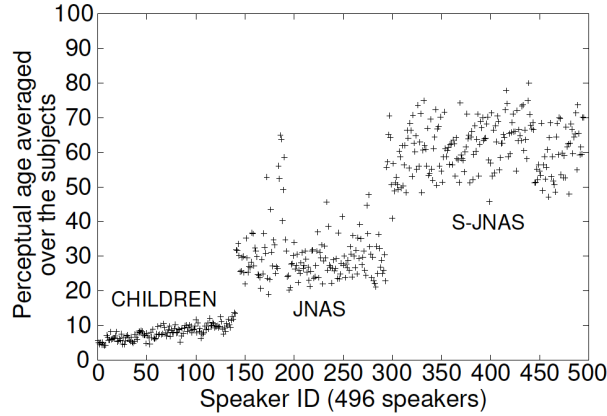


図 5.4: 知覚的年齢の分布

特徴量は MFCC12 次 + Δ MFCC12 次 + Δ パワーの 25 次元である。各話者ごとに GMM16 混合のモデルを作成した。モデル作成用の音声は無音部分を取り除いた 60 秒分の音声が使われた。

入力音声 o が x 歳である尤度は

$$P(x|o) = \frac{P(o|x) P(x)}{P(o)} \quad (5.2)$$

としている。ここで $P(o)$ および $P(x)$ は $P(x|o)$ を最大化する上では定数とみなすことができるので、これを変形して

$$P(o|x) \quad (5.3)$$

の最大化を考える。よって

$$\arg \max_x P() = \arg \max_x P(o|M_x) \quad (5.4)$$

となる。また尤度スコアは入力音声からフレームごとに対数尤度を求め、そのフレーム平均を計算し再び尤度へと戻したものである。

そして、推定された知覚的年齢 PA は以下の式で定義される。

$$PA = \frac{\sum_x x P(o|x)}{\sum_x P(o|x)} \quad (5.5)$$

ただし、この計算に用いられる $P(o|x)$ は尤度値の高いもの上位 11 個である。この 11 個という数字は実験的に得られたもので、最も良い結果が得られたときのものである。

この実験の結果を図 5.5 に示す。知覚的年齢をラベルした聴者は 30 名いるため、ラベルの振り方も 30 種類と平均を取った物を合わせたものの 31 種類が考えられるが、この図は最も相関が高くなったときのものである。そのときの相関は 0.89 であった。またすべてのラベルの種類について相関を求めたものの平均は 0.85 であった。理想的には $y=x$ の直線状に乗ることが期待される。外れた値も少なくないが、おおむねではある程度の範囲に収まっている。

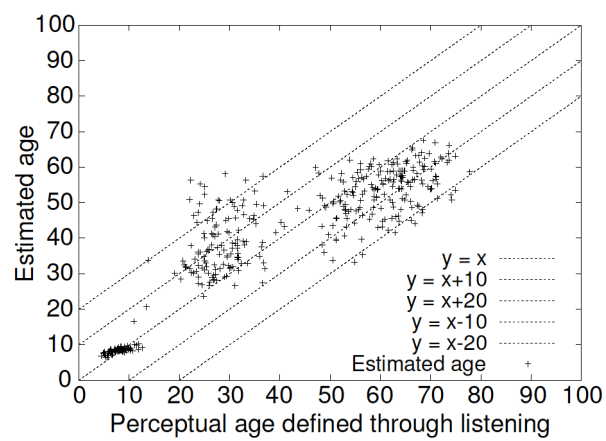


図 5.5: 知覚的年齢とその推定値の相関

第6章 知覚的女声度の聴取実験

6.1 性同一性障害者の音声の収録

まず、女声度推定の実験で用いるための音声を収録を行った。収録した音声は、性同一性障害者 (FtM) の 110 名が発声した音声で、計 140 発声である。ただし、一部の話者は複数回発声している。話者の年齢は 19 歳から 78 歳までひろく分布している。また、発話内容は「ジャックと豆の木」の題名と最初の 1 文で、

「ジャックと豆の木。昔あるところに、ジャックという男の子がいました」

である。この音声には、以降の節で述べるように、聴取実験を行ってどの程度の人に女性として認識してもらえるかというラベルが付与されている。本論文では、これを知覚的女声度と定義する。詳細は以下で述べる。

6.2 知覚的女声度の定義と聴取実験

6.1 節で収録した音声を聴取実験にかけ、女声として受け入れられる割合を求め、ラベルとして付与した。次節以降で行う実験において、計算機の算出した数値の妥当性を示す、あるいは音響的な女声らしさが聴覚的な女声らしさとどのように対応するかを求めるために用いた。

この聴取実験は以下のような条件・手順で行った。

6.2.1 実験条件

実験の意図や目的を伏せた上で、聴者に MtF および生物学的男性・女性の声を聞かせ、話者の性別や年代を判定してもらった。評価者は男性 9 名、女性 9 名の計 18 名である。まず、非 MtF の女性と MtF あわせて 20 人ほどの音声を聞いてもらった後に、実験に用いる 140 の性同一性障害者の音声を評価してもらった。また、これとあわせて、FtM でない女性 15 名の音声も評価者にはわからないように混ぜて評価してもらった。そのほか、評価者には伝えずに、実験の冒頭に男性 2 名、女性 2 名の音声を基準用に評価してもらった。また、各音声について 2 回ずつ評価してもらった。

聴取実験で判定させる内容は

1. 話者の年代（子供,10,20,30,40,50,60,70,80）
2. 話者の性別の度合い

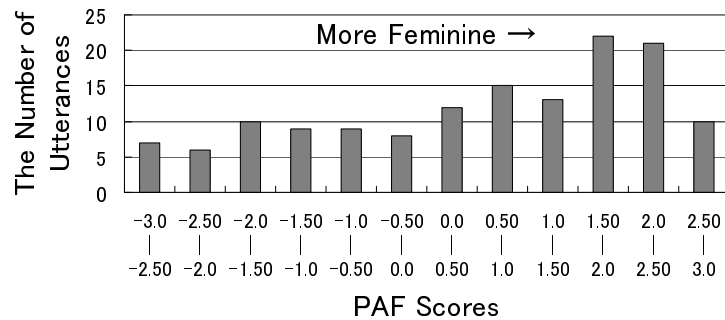


図 6.1: 知覚的女声度のヒストグラム

の2つである。話者の性別の度合については、もっとも男らしいと感じたときには-3、逆にもっとも女性らしいと感じたときには3、どちらでもないときには0と評価してもらった。音声は2秒間隔で1音声ずつ呈示され、1回目が提示された後、話者の順序を変えて再度呈示し、同様の判定をさせた。この聴取実験の結果、各話者について性別の度合いの平均値が算出されるが、この割合を知覚的女声度と定義する。

6.2.2 知覚的女声度の分布

MtFの140話者について、聴取実験によりラベリングされた知覚的女声度のヒストグラムは図6.1のようになった。ただし、各話者の知覚的女声度は2回の聴取結果の平均値である。多少、女声側に偏りがあるものの、良好な広がりを持つことがわかる。以下の実験では、コンピュータによって計算した音響的な女声度と、知覚的女声度の対応を調べるので、女声度には適度な広がりが必要であり、逆にデータに広がりがないと予測などが困難になる。なお、生物学的な女性の音声の知覚的女声度の平均値は2.70であった。

6.3 評価者間の相関と評価者内の相関

前節で調べた聴取結果を用いて、18名の評価者について、評価者内および評価者間の相関係数を調べた。この結果を表6.1に示す。評価者内の相関に関しては、評価者の行った2回の評価値の相関を調べた。評価者間の相関については、各評価者の2回の評価の平均値を求め、その相関を調べた。

また、ある評価者について、その評価者以外の平均値との相関は表6.2のようになった(0.86~0.88)。第8章で述べる女声度自動推定器が出力する推定値と知覚的女声度の相関係数が、表6.2の値と同程度であれば、試作した推定器は十分な性能を持つと言える。

この結果を見る限りでは、男性と女性で声の女らしさの評価を行ううえでの基準には差がみられなかった。

ただし、以降の実験で男女間での評価基準に差が現れることを考え、第7章の実験で用いる知覚的女声度としてまず以下の3つを用いる。

表 6.1: 話者間・話者内の評価値の相関

	中央値	平均値	最小値	最大値
評価者内	0.850	0.833	0.728	0.908
女性間	0.802	0.798	0.692	0.907
男性間	0.784	0.774	0.671	0.887
男女間	0.787	0.781	0.601	0.896
全評価者間	0.787	0.783	0.601	0.907

表 6.2: ある評価者とそのほかの評価者の平均の相関

	中央値	平均値	最小値	最大値
女性	0.884	0.885	0.821	0.932
男性	0.863	0.872	0.824	0.933
全評価者	0.879	0.878	0.821	0.933

1. 男性評価者による知覚的女声度
2. 女性評価者の知覚的女声度
3. 全評価者の知覚的女声度

また、18 人の評価者ごとの知覚的女声度も計算し、一部の実験で用いた。第 7 章以降での実験では、これらの知覚的女声度と、機械により算出した音響的なスコアとの相関係数などを調べた。また、臨床応用での使用時に数字の意味をわかりやすくするため、知覚的女声度の数値を -3 が 0% 、 3 が 100% に対応するように尺度を変えて用いた。この場合、 0% であれば完全に男性の声であることを示し、 100% であれば完全に女性の声として受け入れられるということを意味することになる。

6.4 2 値判定による予備実験

また、この聴取実験より以前に、同様の条件で予備的な実験を行った。基本的な条件は、第 6.2.1 と同じであるが、以下の点が異なる。

1. 評価者は男女計 30 名程度
2. 判定させるのは性別の度合いではなく、男か女かの 2 値判定

この聴取実験によって得られたラベルを使用した場合の実験結果も以降の章で補足として示す。第 6.2 節では、性別の度合いを回答してもらったが、今回は男女の 2 値で、判定してもらい、女性と判定される率を算出した。

第7章 声の女性らしさ推定の枠組み

7.1 既存の技術に基づく孤立母音に対する女声度推定と問題点

これまでの、声質に関する研究では、孤立発声された母音に焦点が当てられている。ここでは、母音の第1および第2フォルマントの周波数を抽出し、母音発声の判別をするといったことが行われてきた。声の女性らしさの推定を考えてみた場合、たとえば孤立母音からは、その音声を発声した話者の声道長を推定することができるが[30]、男性と女性では声道長が異なるため、そこから女声度の推定を行うことが可能ではある。そのほか、声の女らしさが、どのような音響特徴量に現れるかという問題は先行研究でも挙げられている[31][32][10]。

しかし、この技術に基づいて、孤立発声の母音を対象とした女声度推定器を構築し、MtFが注意を払って孤立母音について女性らしく発声できるようになったとしても、それは必ずしも連続発声において女性らしく発声できることを意味しない。これは、たとえるならば、外国語学習において、音素単独では正確に発音ができたとしても、いざ連続的な発声をしようとした場合には、うまく発音できなくなることがある状況と同様である。その原因としては、連続発声における韻律制御の不具合が挙げられるが、しかし、考えられるもうひとつの理由として、連続発声においては各母音の発声に十分注意を払うことができないというものがある。

このように、実際の発話状況を考えると、声の女性らしさの推定器を開発する上では、従来の声の男性らしさ・女性らしさの研究で行われていたような、孤立発声時の母音ごとのフォルマント分析に基づくものは使えない。男女以外の要素としてコンテキストによりフォルマントは変わってくるためである。そのため、連続発話時の音声を対象としたテキスト非依存の技術を用いる必要があることがわかる。

7.2 話者照合技術に基づく声の女性らしさの定義

この問題を解決する方法として、GMMベースのテキスト非依存の話者照合技術を用いることが考えられる。話者認識技術については第5章ですでに詳しく述べた。

テキスト非依存の話者照合技術では、

$$\Lambda = \log P(o|M_s) - \log P(o|M_{\bar{s}}) \quad (7.1)$$

を算出する。ここで、 M_s は話者 s のモデル、 o は観測された音声の話者である。話者照合の判定では、この値をもとにして音声の話者が本人であるかどうかの判定を行う。

本論文の実験では、基本的にこの話者照合の枠組みで声の女性らしさの推定を考える。この声の女性らしさを本論文では音響的女声度 (AF: Acoustic Femininity) と呼び、これを以下の式で定義する。

$$AF(o) = \log P(o|M_F) - \log P(o|M_M) \quad (7.2)$$

ここで、 M_M 、 M_F はそれぞれ男声モデル、女声モデルを表す。

7.3 女性らしさの程度を算出する妥当性

話者照合では計算によって得られた値が決められた閾値より大きければ、音声を発した人が、その主張する人物であると受諾し、そうでなければ拒否する。そこでは、入力した音声はどの程度その人物らしいかという情報は失われ、受諾か拒否かという2カテゴリーのカテゴリ分けの結果のみが残る。

次に、本論文のような音響的女声度の推定を考えてみる。確かに、性同一性障害者ではない人物であれば、生物学的な女性であれば、完全に女性の声、生物学的に男性であれば完全に男性の声と分かれる。参考文献 [23] では、コンピュータによって音声から男女を識別する場合に 100% の正解率が得られることが示されている。

性同一性障害者は、特に MtF の人物の場合、生物学的には男性でありながら、女性的な声を出すことを望み、そのような発声を行おうとする。まだ完全に男性的な音声を発声するものから、男性とも女性ともいえない中間的な声を出すもの、なかには完全に女性と聞こえる声を出せる人物が存在する。そのため、声の男性らしさ・女性らしさは幅広く分布することになる。

そこで、本論文の実験では、入力音声に対し男声か女声かの判定をすることではなく、どの程度の女声らしさを持つかの程度を算出することを最初の目的とした。参考文献 [21] では、人間による声の性別・男女性の知覚は、連続的・段階的であり、カテゴリー的ではないという報告がされている。このため、コンピュータによる声の女性らしさの推定においても、程度 (degree) を算出することは妥当であると考えた。

7.4 音響特徴量の検討

7.4.1 ボイスセラピーとの関連からの検討

声の女性らしさを推定するうえで、どのような音響特徴量を用いるかということは、すなわち、声の女性らしさがどのような特徴量に集中して現れるかという問題である。第2章でも述べたが、ボイスセラピーにおいて、男性の声を女性化するときには、以下の三つの点に注目する。

1. 声の高さ (基本周波数)
2. 喉の絞り具合 (すなわち、声道の形状の情報に相当)
3. 女性らしい話し方 (抑揚の付け方など)

1つ目の声の高さに関しては、女性は男性よりも声が高いため、女性らしく聞こえる声を出すためには、より高い声をだす必要があるということである。これは第3のソース

フィルタモデルにおける韻律的特徴の一部であり、音響特徴量としては基本周波数 (F_0) として表される。音声全体にわたる声の高さの事を指しているため (ピッチレンジ)、静的な情報のみで動的な情報は基本的に含まれないことになる。

2 つ目の喉の絞り具合は、声道形状を変化させることを意味している。男性と女性では声道の長さや太さが異なり、男性のほうが 1.2 倍ほどサイズが大きい。より女性らしい声になるように、のどを絞る、すなわち声道を狭めたり、声帯を持ち上げるといったことをする必要があるのである。これを表す音響特徴量としては、ソースフィルタモデルにおける分節的特徴に相当する、スペクトル包絡に関連したものが挙げられる。

3 つめの話し方であるが、これはたとえば声の高さの抑揚 (ピッチパターン) に相当するものである。そのほかにも、声の出だしの柔らかさや硬さ (パワーの変化) の違いなども含まれる。この情報を取り扱う手法はまだ確立されているとはいいがたい。例えば、音声合成における韻律制御はまだ十分な清野が得られていない。また、第 4 節でもとりあげたが、多くの研究では、基本周波数やパワーについて発話全体に対する統計量 (平均や分散など) を計算し、あとは機械学習による識別に任せるかたちをとっていることから、そのことが分かる。また、発話内容に対する依存性も高くなる。このことは、発話内容に対する依存性を持たないシステム作りを目的とした本論文の実験とも相性が悪いということになる。

7.4.2 使用する音響特徴量

そこで、本論文の実験では、パラメータとしては用いる音響特徴量は、声道形状に対応しスペクトル包絡の特性を表現する特徴としては MFCC を、声の高さに関する静的な情報に対応する特徴としては $\log F_0$ を採用した。韻律的特徴として声帯波形の男女差によるスペクトル包絡への影響も指摘されているが [17][19]、参考文献 [21] では、声帯波形の形状の差が生み出すスペクトル包絡の相違が声の女性らしさの知覚にたいして有意には貢献していないという報告がある。本論文ではこれにならい、声帯波形の違いに起因する要素に関する特別な考慮は行わない。

7.5 GMM ベースのモデルの構築

そのほかの条件を合わせて、表 7.1 に実験条件をまとめる。実験に用いる 2 つの音響特徴量それぞれについて、男声のモデルと女声のモデルを構築した。つまり、MFCC を音響特徴量とする男声のモデルと女声のモデル、 $\log F_0$ を音響特徴量とする男声のモデルと女声のモデルの合計 4 つのモデルを作成した。また、各モデルは 16 混合の GMM とした。以下、音響的女声度については、 AF^s はスペクトル (MFCC)、 AF^f は $\log F_0$ に関するモデルに基づく音響的女声度をあらわす。これを式で表すと次のようになる。

F_0 推定は Praat[33] を用いて自己相関法で行った。音声はまずパラメータ系列に変換し、そののちに、学習用・評価用音声の双方に対して無音区間除去の処理を施した。また今回の実験では音声における男女間の差が問題となるが、その話者性の違いが現れるのは主に母音の部分であるため、実験に必要な子音の区間の除去も行った。以上の操作はパワーに着目した処理で行った。具体的には、平均パワーよりもパワーの低い部分が 50 (ms) 以

表 7.1: 分析条件

サンプリング	16bit / 16kHz
窓	窓長 25msec、シフト長 10msec
パラメータ	MFCC12 次元+ Δ 12 次元+ Δ E 計 25 次元 $\log F_0$

上続いている部分に対応するフレームを除去することで行った。ただし、平均パワーは、最大パワーの 0.5 倍から 0.95 倍の部分のみから求めた。また、 F_0 の抽出できなかったフレームも除去した。

$\log F_0$ についても、単純に $\log F_0$ を用いた閾値処理をするのではなく、このようなモデル化をしてそのモデルに対する対数尤度を用いることで、たとえば通常の女性と比べても極端に高い F_0 をもつような発声をした場合でも、スコアは必ずしも高くはないことが期待できる。

$$AF^s(o) = \log P(o|M_F^s) - \log P(o|M_M^s) \quad (7.3)$$

$$AF^f(o) = \log P(o|M_F^f) - \log P(o|M_M^f) \quad (7.4)$$

7.6 線形回帰による拡張と予測値の算出

次に、話者照合に基づいた推定をより一般に拡張し、線形回帰分析により、入力音声聴取実験において女性と判定される率の予測値の算出を行った。線形回帰予測による女声度推定システムの概要図を図 7.1 に示す。このときの説明変数としては、MFCC、 $\log F_0$ についてそれぞれの男声モデルおよび女声モデルの対数尤度、あわせて 4 つの対数尤度のすべてを用いた。

第 7.5 節においては式 7.3 および式 7.4 を定義した。そこでは、各モデルの対数尤度に対する係数は 1 に固定していた。線形回帰による予測では、これを一般に拡張した次の式を用いる。

$$\begin{aligned} AF(o) = & \alpha \log P(o|M_F^s) + \beta \log P(o|M_M^s) \\ & + \gamma \log P(o|M_F^f) + \epsilon \log P(o|M_M^f) \\ & + C \end{aligned} \quad (7.5)$$

ただし各モデルに対する重み係数は最適化されることになる。本論文では、目的変数は聴取実験によって得られた知覚的女声度とし、最小二乗法により、予測残差の二乗和が最も小さくなるように係数の最適化を行った。

7.7 実験に用いたコーパス

実験に用いたコーパスは以下の 2 つである。

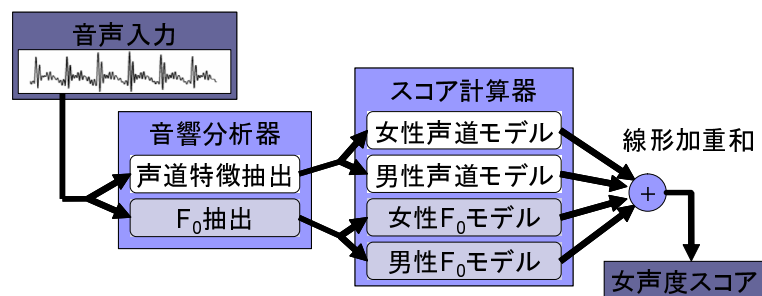


図 7.1: システムの概観

1. 新聞記事読み上げ音声コーパス (JNAS)

2. 性同一性障害者の音声

性同一性障害者の音声は第 6 章で収録し、聴取実験に用いたものと同一である。以下、JNAS の音声データベースについて詳しく述べる。

7.7.1 新聞記事読み上げ音声コーパス (JNAS)

新聞記事読み上げ音声コーパスは、日本語の大語彙連続音声認識の研究に用いる目的で作られたコーパスである [34]。コーパスの詳細は表 7.2 に示すとおりである。男女それぞれ約 150 名が、それぞれ約 150 文を発声しているが、そのうち用いたのは男女各 114 名、各話者 30 発声で計 6840 発声ある。これは計算機で学習する上での制約によるものであり、理想的にはすべてのデータを用いるのが望ましい。

表 7.2: 新聞記事読み上げ音声コーパスの概要 [34]

話者		男女各 153 名 (計 306 名)
読み上げテキスト	新聞記事文	155 セット (約 100 文/セット, 計 16,176 文)
	音素バランス文	10 セット (約 50 文/セット, 計 503 文)
文数/話者	新聞記事文	1 セット
	音素バランス文	1 セット
総発話数	新聞記事文	31,938 発話
	音素バランス文	15,372 発話
録音時間 (新聞記事文)		215,247 秒 (約 59 時間 47 分)
収録サイト		39 機関
収録マイク		headset と desktop の 2 本で 2ch 収録
音声データ	A/D	16bit 量子化, 16kHz サンプルング
	ヘッダ	NIST Sphere
	圧縮	Shorten

第8章 実験結果

8.1 単純な推定の結果

第7章ので作成した男女の MFCC および F_0 をパラメータとする GMM モデルに対し、MtF 話者の音声を入力して音響的女声度を算出した。このときの、MtF 話者の音声の音響的女声度 (A^s 、 A^f) と、第6章で求めた知覚的女声度の相関は表 8.1 のようになった。知覚的女声度としては女性評価者平均、男性評価者平均および全評価者平均から得られたものを示した。このうち、全評価者平均の知覚的女声度を用いた場合のの散布図は、図 8.1、8.2 のようになった。それぞれ、パラメータとして用いる音響特徴量が MFCC、 $\log F_0$ のときである。より女性らしい声を出すためには声の高さだけではなく、声道形状も重要であることが示された。

8.1.1 議論

今回、入力用に用いた音声の中には、 F_0 のみを不自然に高く上昇させたものも多数含まれており、それらの音声に対して適切な評価をするかが問題となる。音源情報のみで評価した場合は先に述べたような、 F_0 のみを高くした音声でも、 F_0 が女性のレンジに含まれている場合は、実際よりも女声度がかかなり高く判定されてしまう。これは図 8.2 で直線より右下に外れている音声に対応する。このような極端な音声に対しては MFCC をパラメータとして使用することである程度は適切に判定できるが、依然として回帰直線から乖離する音声が存在した。パラメータとしては F_0 、MFCC のどちらも、単独で用いた場合では女声度推定には限界がある。

8.2 線形回帰による予測値の算出

次に、線形回帰分析により、入力音声が聴取実験において女性と判定される率の予測値の算出を行った。予測値と実際の知覚的女声度の相関係数は表 8.2 のようになった。また、散布図を図 8.3、8.4、8.5 に示す。それぞれ、ラベルが全評価者平均、男性平均、女性平均のときである。このとき、最後の線形回帰による予測では評価データが学習データに含まれているためクローズドな条件になっている。

予測するというタスクについて、未知データに対する性能を見るために、予測器の学習において評価データを学習データに含めない発話オープンな条件下でも実験を行った。そのときの相関係数は表 8.3 のようになった。全評価者平均の結果を第 8.1 節の結果と比べることで、MFCC の表す声道形状の情報と、音源情報をあわせて用いることにより、高い相関を得ることができたということがわかる。

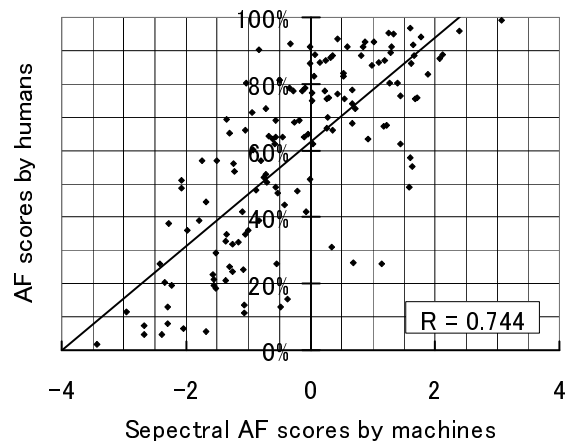


図 8.1: 声道特性に基づく音響的女声度と聴取結果による知覚的女声度の関係

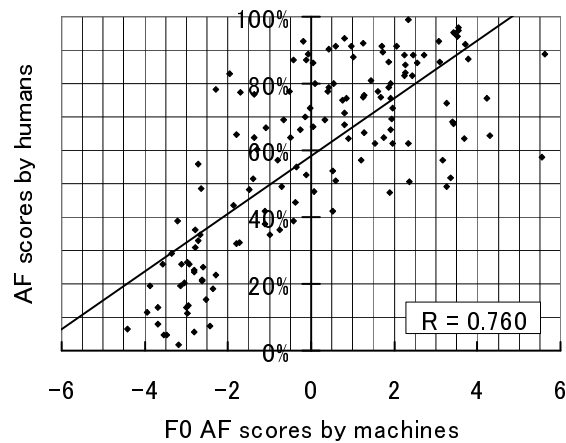


図 8.2: 音源特性に基づく音響的女声度と聴取結果による知覚的女声度の関係

この知覚的女声度推定の精度としては、第 6 章において、ある評価者のスコアとその他の全評価者の平均スコアの相関係数が 0.88 程度である（表 6.2）ことを考えると、0.86 という相関係数は十分な性能を示しているといえる。

また、第 6 章の聴取実験では、性同一性障害者以外にも成人女性の音声も同時に聴取実験にかけた。そこで、この成人の生物学的女性の音声（15 人、各人 1 発声）を入力として用いて同様に知覚的女声度の推定を行った。このとき、聴取実験による実際の知覚的女声度では 91.3% であるのに対し、コンピュータによる知覚的女声度の予測値は 86.8% となった。これにより、MtF でない話者の音声に対しても十分な精度を持つことが分かった。

表 8.1: 音響的女声度と知覚的女声度の相関

	female PAF	male PAF	avg. PAF
AF_s	0.735	0.737	0.743
AF_f	0.756	0.748	0.760

表 8.2: 知覚的女声度とその予測値の相関

	female PAF	male PAF	avg. PAF
AF	0.848	0.866	0.863

表 8.3: 発話オープンな条件下での知覚的女声度とその予測値の相関

	female PAF	male PAF	avg. PAF
AF	0.836	0.854	0.853

8.2.1 線形回帰係数

線形回帰方程式の係数を表 8.4 に示す。これをみると、スペクトルに関しては男女各モデルに対する対数尤度に乗算される係数の絶対値はほぼ同じである。つまり、女性に近いことと男性から遠いことが同程度に評価されるということである。一方 F_0 に関しては男性モデルに対する対数尤度に乗算される係数は、女性モデルのそれに比べて非常に小さい。特に、評価者平均のラベルに関してはほぼ 0 とみなしてよい。 F_0 に関しても GMM によるモデル化をし、そのモデルに対する対数尤度を計算していることを考えると、この結果は声の高さは女性として自然な範囲に収まっていけば高くても低くても同じように評価されるということを意味する。すなわち、MtF は生物学的に男性であり、そのため声帯を支える筋が重いわけであるが、極端に高い声を出す必要はないということである。

線形回帰係数から男性平均と女性平均について各係数を見てみると、女性は男性と比較して、スペクトルの情報を重視する傾向があるということが分かる。そこで、 α 及び β に対する有意差検定（分散分析）を評価者の男女間で行なった。その結果、 α, β いずれにおいても、男女間の分布の分離は 25% の危険率として算出された。女性評価者が、男性と比較して、声道形状に起因する音響量を統計的に有意に重要視していると強く主張することはできないが、表 8.4 は、女性が声道形状を重視する傾向にあることを示す結果であると解釈している。

男性に女性と認識してもらいたい場合と、女性に女性として認識してもらいたい場合で基準が異なるということも言える。女性評価者の場合、線形回帰による重相関係数も男性と比較し若干低い。これは、今回評価に用いた MFCC（声道特性）や F_0 （音源特性）の静的な情報だけでは、女性による評価は説明がしづらいことを示唆する。評価者が女性の場合、声の抑揚が女性らしいかといったダイナミクスに関する部分も評価に影響している可能性も考えられる。

表 8.4: 線形回帰係数

	α M_F^s	β M_M^s	γ M_F^f	ϵ M_M^f	C
fem. PAF	0.097	-0.092	0.121	-0.012	1.071
male PAF	0.080	-0.075	0.122	0.012	1.075
avg. PAF	0.089	-0.084	0.121	-0.000	1.073

8.3 知覚的女声度のラベルとして 2 値の聴取結果を用いた場合

ラベルとて、聴取実験において 2 値による評価をさせたときに女声と判定される率（女声判定率）を用いた場合の結果を示す。このとき、GID 話者の音声の音響的女声度と女性判定率の散布図は、MFCC、 $\log F_0$ についてそれぞれ図 8.6、8.7 のようになった。また、音響的女声度と女性判定率の相関係数はそれぞれ 0.717、0.704 となった。

8.3.1 予測値の算出

次に同様にして、線形回帰分析により、女声判定率の予測値を算出した。このときの散布図を図 8.8 に示す。相関係数は 0.799 となった。ラベルとして性別の度合いを評価させたもの（知覚的女声度）を用いた場合と同様に、MFCC の表す声道形状の情報と、音源情報をあわせて用いることにより、高い相関を得ることができた。知覚的女声度をラベルとして用いた場合に比べて、音響的女声度との相関は低くなっている。これは、聴取実験において 2 値による評価をさせたことにより、カテゴリー知覚的な判断がなされたために、女声判定率が 50% 付近の音声に対する評価にばらつきが生まれたためと考えられる。

8.3.2 GID 話者の音声によるモデル

前節までは JNAS のデータからモデルを作成したが、それとは別に、GID 話者の音声を用いて、男女のモデルを作成することもできる。すなわち、女性判定率の低い音声から男性モデル、女性判定率の高い音声から女性モデルを作成するということである。これは、MtF の話者にとっては、ボイスセラピーを受けたことによって、女性判定率の高くなった声こそ目指すべき声であるといえるからである。また、逆に本人は女性の声として発声していながら、多くの人に男性と判定される音声からは遠い方が望ましい。

そこで、データベースとして JNAS 音声を用いた場合と同様にして、かわりに GID 話者の音声を学習用音声として用い、GID 話者の男性モデルと女性モデルを作成した。ここで、男性モデルの学習データには女性判定率が 60% 以下の音声、女性モデルの学習データには女性判定率が 60% 以上のものを使用した。ただし、評価用音声と同一話者の音声は学習用データから除外した。また音響パラメータは MFCC のみを試した。このときの女声度と女性判定率の相関係数は 0.749 となった。JNAS の音声からモデルを作成した場合に比べ、高い相関が得られた。

つぎに、第 8.3.1 節と同様にして線形回帰分析を行った。ここで説明変数としては、第 8.3.1 節で用いた 4 つの対数尤度に、GID 話者モデルの対数尤度 2 つを加えた 6 つの対数

尤度を用い、重線形回帰分析を行い、予測値を算出した。この予測値と女性判定率の相関係数は 0.807 となった。GID 話者モデルを合わせることで、相関はわずかながら向上したが、特筆するほどの効果は得られなかった。

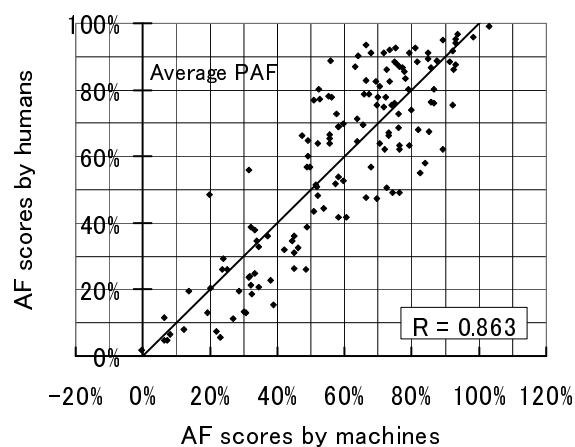


図 8.3: 知覚的女声度の予測値と知覚的女声度の関係 (全評価者平均)

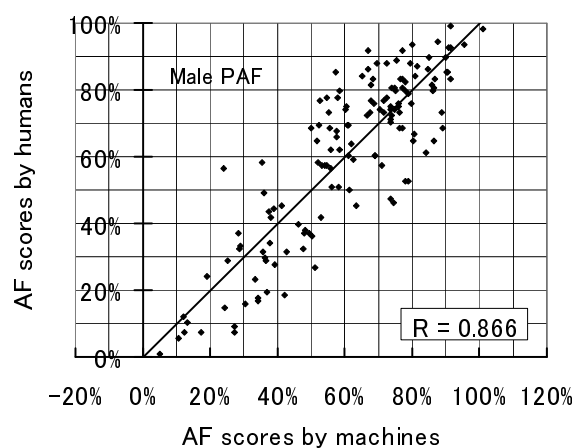


図 8.4: 知覚的女声度の予測値と知覚的女声度の関係 (男性評価者平均)

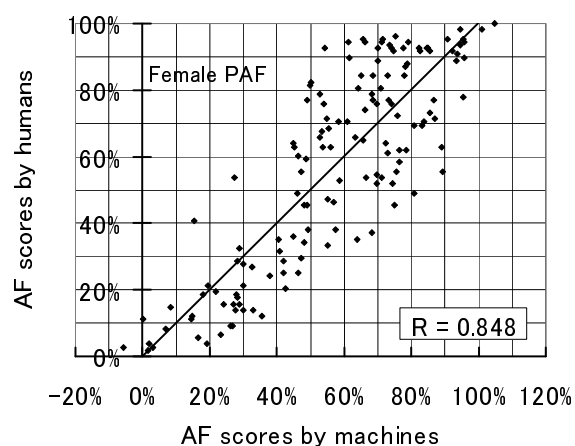


図 8.5: 知覚的女声度の予測値と知覚的女声度関係 (女性評価者平均)

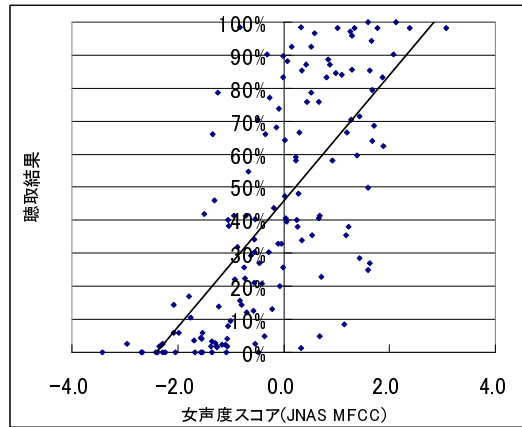


図 8.6: MFCC を音響特徴量として用いた場合の音響的女声度と 2 値評価による聴取結果の相関

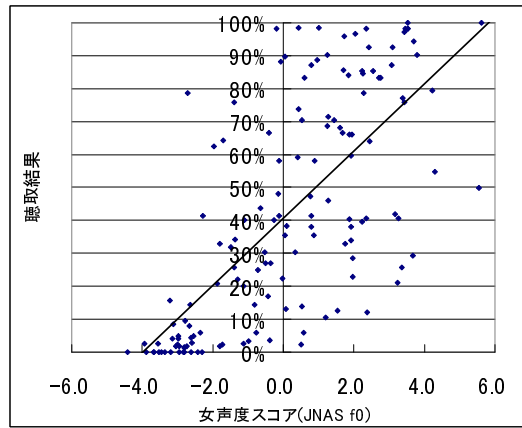


図 8.7: $\log F_0$ を音響特徴量として用いた場合の音響的女声度と 2 値評価による聴取結果の相関

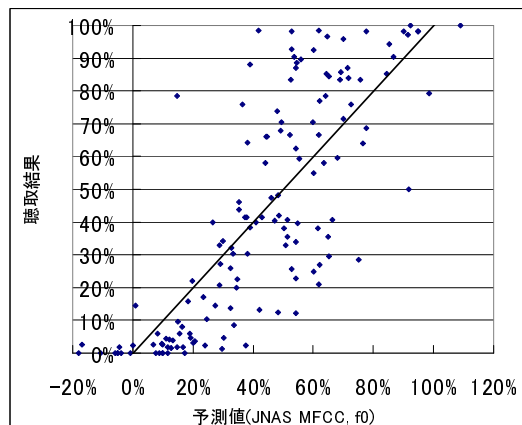


図 8.8: 線形回帰による予測と 2 値評価による聴取結果の相関

第9章 女声度推定器のインターフェース化と臨床の場面への導入

9.1 インターフェースの実装

第8章まで述べた知覚的女声度推定器にインターフェースを実装し簡単に使用できるようにした。これを図9.1に示す。ここでは、ボイスセラピーの前後で声が女性化された人物の、ボイスセラピーの前と後の音声に対する知覚的女声度の推定結果が示されている。

9.2 臨床応用

櫻庭氏のボイスセラピーに、2006年2月より随時導入していただいている（図9.2、図9.3）。また、要望などのフィードバックをいただき、必要に応じてインターフェースの改良などを行っている。ボイスセラピーでは、MtF当事者が実際に音声を入力し、その声に対する知覚的女声度の推定結果を見ていただき、それに対して推定器がどのような反応を見せるかを観察した。

その結果、以下のような傾向や実用性があることがわかった。

1. ボイスセラピーにおいて、セラピストが初回にMtFの声を評価すると、ほとんどの場合に男声と判定されることになる。そのときに、MtF当事者にそのMtFの声が「男声」であると告げる必要があるが、セラピストの聴覚印象だけで判定結果を告げるよりも、機械を使うことで、実際に数値で表示される方がMtF当事者に判定結果伝わりやすくなる。
2. ボイスセラピーの第一段階では、まず声の高さのみに着目し、高い声を出す練習を行う。知覚的女声度推定器では、高い声で発声すると女声として不自然にならない範囲で女声度が上がる。そのため、ボイスセラピーの第一段階で用いるのに有効である。
3. ボイスセラピーを行うときに、その開始前と終了後で知覚的女声度を測定する。ボイスセラピーにおける発声練習の前と発声練習の後では、多くの場合、知覚的女声度の推定値が向上する。そのため、発声練習が大切であることをMtF当事者に実感していただける。

しかしながら、同時に以下のような改良すべき点も見受けられた。

1. 知覚的女声度推定器では、声道を狭めて出した高い声には高得点が出やすい傾向があり、話し方が男らしくとも、女声と判定されてしまう。これは、今回構築した知



(a) 女性化する前の声



(b) 女声化後の声

図 9.1: 知覚的女声度推定器のインターフェース



図 9.2: 知覚的女声度推定器の実際の使用の様子

覚的女声度推定器では、韻律的特徴に関しては、声の高さの平均的な情報のみを用いており、そのパターンを見ていないことや、

2. 逆に話し方が女性らしくても、声が低いと女声度が低く抑えられる傾向がある。
3. 以上の点から、一部の音声では、第三者による聴取実験から得られる知覚的女声度と知覚的女声度推定器による推定値に乖離がみられ、(1) \ (2) のような音声に関しては、適宜、第三者による聴取実験を行う必要性がある。
4. 入力とする音声が単独の母音であるばあい、知覚的女声度推定器による推定値が高くなる傾向がある。喉をしぼる、あるいはつめるようにして発声し、高い声を出しやすい母音では、推定の結果が 140% を超えることもありえる。生物学的な女性の自然な発声を入力とした場合の推定値よりも、はるかに高い推定値が算出されるこ



図 9.3: 知覚的女声度推定器の実際の使用における画面表示

とがあるということである。

5. 第三者による聴取実験では、MtF の発声において、話し方が男性的であると、声の高さを上げてもなかなか女声であると判定されない傾向も見られるが（カテゴリー知覚的）、知覚的女声度推定器では、声の高さに対して直線的に推定値が上昇する。

現在、実際の臨床の場面で知覚的女声度推定器を使用していただく際には、この推定器で判定できること、および推定器の限界や制限を事前に MtF 当事者に説明し、十分な理解を得た上で使用するようにしていただいている。発声全体にわたる声の高さや、喉をしぼるように発声する変化した声道形状に起因する女声らしさは、今回構築した推定器でも十分に評価が可能であるが、一方で女声らしいイントネーションややわらかい声の出し方といった部分については、セラピストの判断と指導にゆだねるということである。

現段階は知覚的女声度推定器には、セラピストから見た場合に改良すべき点を多く含むものの、発した音声の客観的な女声度を数値として見ることができるため、MtF にとって訓練の目安になり、また励みにもなっているという評価をいただいた。

第10章 結論

性同一性障害者、特に MtF を対象として、その音声がどの程度女性らしく聞こえるかをコンピュータにより自動的に推定する手法を提案した。これはテキスト非依存の話者照合技術に基づいており、発話内容を問わず、連続的な音声に対して音響的な女声度を評価することができるものである。同時に、聴取実験を行い第三者に MtF の音声に対してどの程度女声らしく聞こえるかという判断を下してもらい、人間の判断する声の女性らしさ、すなわち知覚的女声度を定義し、MtF の音声に対するラベルとした。コンピュータによる音響的女声度の推定結果と、聴取実験による知覚的女声度を比較することによって、より女性らしく聞こえる声で話すためには、ただ声を高くするだけでなく、声道形状を適切に制御することもある必要であるということが分かった。

また臨床応用も考え、線形回帰分析を用いて聴取実験によって得られる知覚的女声度の予測を行った。予測値と実際の聴取結果との相関係数は、0.86 と良好な結果を得ることができた。これと対応する人間同士の評価の相関係数の平均が 0.88 であったことから、この結果は、知覚的女声度自動推定器をひとりの人間とみなせることを示している。

現在の知覚的女声度推定器では、話し方の女性らしさについては考慮していないが、発話の男性らしさ、女性らしさに、話し方の要素の影響がほとんどないということは考えづらい。今回の実験では MtF の発声した音声資料は文章朗読であったため、話し方の違いが影響としてあらわれにくいとも考えられる。すなわち、聴取実験で得られた知覚的女声度にも、話し方の影響が現れておらず、その結果、知覚的女声度を予測するタスクが容易になったため、高い精度が得られたということもありうる。

女声らしい話し方についても検討し、判定に加えることは今後の課題となる。そのためには適した音声の収録などが必要になるであろう。

臨床応用としては、すでに 2006 年 2 月から TVT に実験的に随時導入しており、改良の余地はあるものの、声の高さを上げて話す訓練の時には有効であり、また練習の成果が数値化されることで、一般的に訓練の励みになることがわかった。

謝辞

広瀬先生と峯松先生には、卒論のときより3年にわたり、日ごろから多大な指導と恩顧を頂きましたことを、この場をお借りして感謝いたします。また、櫻庭様には、性同一性障害やボイスセラピーに関するさまざまな知見を頂き、本論文を完成させることができました。ここにお礼を述べさせていただきます。

また、職員の高橋様をはじめとして、秘書の武田様、笠島様、広瀬・峯松研究室の皆様を支えられて、研究生生活を送ることができました。ここに感謝いたします。

参考文献

- [1] 野宮亜紀他, “性同一性障害って何”, 緑風書店, 東京 (2003)
- [2] DSM-IV-TR 精神疾患の分類と診断の手引 新訂版, 米国精神医学会, 高橋三郎他訳, 医学書院, 東京 (2003)
- [3] M. T. Edgerton, “The surgical treatment of male transsexuals,” *Clinics in Plastic Surgery*, 1(2), pp.285–323 (1974)
- [4] F. G. Wolford and R. G. Parry, “Laryngeal chondroplasty for appearance,” *Plastic and reconstructive surgery*, 56(4), pp.371–374 (1975)
- [5] R. C. Bralley, G. L. Bull, C. H. Gore, and M. T. Edgerton, “Evaluation of vocal pitch in male transsexuals,” *J. Communication Disorder*, 11, pp.443–449 (1978)
- [6] L. E. Spencer, “Speech characteristics of MtF transsexuals: a perceptual and acoustic study,” *Folia phoniat.*, 40, pp.31–42 (1988)
- [7] K. H. Mount and S. J. Salmon, “Changing the vocal characteristics of a postoperative transsexual patient: a longitudinal study,” *J. Communication Disorder*, 21, pp.229–238 (1988)
- [8] 櫻庭京子, 今泉敏, 広瀬啓吉, 新美成二, 箕一彦, “女性と判定された性同一性障害者 (MtF) の声の基本周波数”, 電子情報通信学会音声研究会, sp2002-187, Vol.102 No.749 (2003)
- [9] 櫻庭京子, 今泉敏, 広瀬啓吉, 新美成二, 箕一彦, “女声と聴取された性同一性障害者 (MtF) の音声の音響分析”, 日本音響学会春季講演論文集, pp.449-450 (2003)
- [10] M.P. Gelfer, K.J.Schofield, “Comparison of acoustic and perceptual measures of voice in male-to-female transsexuals perceived as female vs. those perceived as male,” *J. Voice*, 14, pp.22-33 (2000)
- [11] Wolfe, V.I., “Intonation and fundamental frequency in MtF TS,” *J. Speech Hearing Disorders*, 55, pp.43-50 (1990)
- [12] 新美成二, 田山二郎, 今泉敏, 山口宏也, “音声生成の科学—発声とその障害—”, 医歯薬出版株式会社, 東京 (2003)
- [13] S.McGilloway, R.Cowie, E.Douglas-Cowie, “APPROACHING AUTOMATIC RECOGNITION OF EMOTION FROM VOICE: A ROUGH BENCHMARK,” *Proc. ISCA Workshop Speech and Emotion*, pp.207-212 (2000)

- [14] I.Luengo, E.Navas, I.Hernández, J.Sánchez, “Automatic Emotion Recognition using Prosodic Parameters,” Proc. EUROSPEECH 2005, pp.493-496
- [15] D.Ververidis, C.Kotropoulos, I.Pitas, “AUTOMATIC EMOTIONAL SPEECH CLASSIFICATION,” Proc. ICASSP, pp.593-596 (2004)
- [16] J.Cichosz, K. Ślot, ”Low-Dimensional Feature Space Derivation for Emotion Recognition,” Proc. EUROSPEECH, pp.477-479 (2005)
- [17] 梅田規子, 寺西立年, “声の韻質と声質”, 日本音響学会誌, Vol.22, No.4, pp.195–203 (1966)
- [18] 佐藤大和, “女声を特徴づける音響パラメータの研究”, 電子通信学会論文誌, Vol.57-A, No.1, pp.23–30 (1974)
- [19] 安広輝夫, 尾関和彦, “男声・女声変換実験”, 日本音響学会誌, Vol.32, No.6, pp.362–368 (1976)
- [20] 太田耕三, “推定された母音声道面積関数の多変量解析”, 日本音響学会誌, Vol.34, No.11, pp.624–634 (1978)
- [21] John W. Mullenix, Keith A. Johnson, Meral Topcu-Durgun, Lynn M. Farmsworth, “The perceptual representation of voice gender,” J. Acoust. Soc. Am., Vol.98, No.6, pp.3080–3095 (1995)
- [22] Ke Wu, D. G. Childers, “Gender recognition from speech. Part I: Coarse analysis,” J. Acoust. Soc. Am., Vol.90, No.4, pp.1828–1840 (1991)
- [23] D. G. Childers, Ke Wu, “Gender recognition from speech. Part II: Fine analysis,” J. Acoust. Soc. Am., Vol.90, No.4, pp.1841–1856 (1991)
- [24] J.P. Campbell, “Speaker Recognition: A Tutorial,” Proc. IEEE, Vol.85, No.9, pp.1437–1462 (1997)
- [25] 北脇信彦, 菅村昇, 小泉宣夫 “音のコミュニケーション工学,” コロナ社 (1996)
- [26] A. E. Rosenberg and S. Parthasarathy, “Speaker background models for connected digit password speaker verification,” Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, pp.81–84 (1996)
- [27] L. P. Heck and M. Weintraub, “Handset-dependent background models for robust text-independent speaker recognition,” Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, pp.1071–1074 (1997)
- [28] Z.Lei, Y.Yang, Z.Wu, “Mixture of Support Vector Machines for Text-independent Speaker Recognition,” Proc. EUROSPEECH, pp.2041-2044 (2005)

- [29] 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.2186-219, (2002)
- [30] A. Paige *et al.*, “Calculation of vocal tract length,” IEEE Trans. on Audio and Electroacoustics, Vol.AU-18, No.3, pp.268-270 (1970)
- [31] S. Bennett, “Acoustic correlates of perceived sexual identity in preadolescent children’s voices,” J. Acoust. Soc. Am., Vol.66, No.4, pp.989-1000 (1979)
- [32] M. L. Andrews et al., “Gender presentation: perceptual and acoustical analyses of voice,” J. Voice, Vol.11, No.3, pp.307-313 (1997)
- [33] Praat: doing phonetics by computer,
(<http://www.fon.hum.uva.nl/praat/>)
- [34] JNAS: Japanese Newspaper Article Sentences,
<http://www.mibel.cs.tsukuba.ac.jp/jnas/>

発表文献

- [1] 丸山和孝, 櫻庭京子, 峯松信明, 広瀬啓吉, 田山二郎, 今泉敏, 山内俊雄, “話者認識技術を用いた性同一性者の音声に対する男声度・女声度の自動推定”, 日本音響学会春季講演論文集, 3-P-20, pp.489-490 (2006-3)
- [2] 櫻庭京子, 丸山和孝, 峯松信明, 広瀬啓吉, 山内俊雄, 田山二郎, 今泉敏, “男性から女性への性別の移行を希望する性同一障害患者 (MtF) の音声認識による評価と臨床応用”, 日本音響学会春季講演論文集, 3-P-21, pp.491-492 (2006-3)
- [3] 櫻庭京子, 丸山和孝, 峯松信明, 広瀬啓吉, 田山二郎, 今泉敏, 山内俊雄, “Transsexual voice therapy における話者性別推定技術の臨床応用”, 性同一性障害研究会 (2006-3)
- [4] 櫻庭京子, 丸山和孝, 峯松信明, 広瀬啓吉, 田山二郎, 今泉敏, 山内俊雄, “性同一性障害者 (MtF) の音声に対する知覚的性別の自動推定”, 電子情報通信学会音声研究会, SP2005-189, pp.29-34 (2006-3)
- [5] 丸山和孝, 櫻庭京子, 峯松信明, 広瀬啓吉, 田山二郎, 今泉敏, 山内俊雄, “話者照合技術に基づく性同一性障害者の音声に対する女声度の自動推定”, 日本音響学会秋季講演論文集, 3-P-13, pp.361-362 (2006-9)
- [6] N. Minematsu, K. Maruyama, K. Sakuraba, K. Hirose, N. Tayama, S. Imaizumi, and T. Yamauchi, “Development of a femininity estimator using speaker recognition techniques for voice therapy of gender identity disorder clients,” Proc. Int. Conf. Acoustics, Speech, & Signal Processing (ICASSP'2007) (2007-4, accepted)