社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE

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

櫻庭京子[i]，丸山和孝[ii]，峯松信明[iii]，広瀬啓吉[ii]，田山二朗[iv]，今泉敏[v]，山内俊雄[vi]

i 清瀬市障害者福祉センター　ii 東大・情報理工　iii 東大・新領域

iv 国立国際医療センター　v 県立広島大　vi 埼玉医大

E-mail:　sakuraba@mtd.biglobe.ne.jp

あらまし　著者ら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 transgendered/transsexual = MtF）に対して、声を女性化させるための transsexual voice therapy を行っており、今回の発表では話者認識技術を用いて知覚的女性度を推定するシステムの臨床応用について検討する。このシステムは、声道特性と音源特性それぞれについて、男声モデル・女声モデルを持ち、各特性別に入力音声の女声度を推定、聴取実験により女性と判定される率（知覚的女声度）の予測値を算出する。上記のシステムを実際の臨床で用いた結果、声道形状を変えながらピッチをあげて女声をつくる方略の完成度を知ることはできるが、発話スタイルの動的制御に基づく女声の生成方略には対応できておらず、今後の検討の課題であることがわかった。

キーワード　男女識別、自動推定、話者認識、性同一性障害、transsexual voice therapy

Automatic estimation of femininity of MtF's speech using speaker recognition technique and its clinical application

Kyoko SAKURABA[i], Kazutaka MARUYAMA[ii], Nobuaki MINEMATSU[iii], Keikichi HIROSE[ii]

Niro TAYAMA[iv], Satoshi IMAIZUMI[v], Toshio YAMAUCHI[vi]

i Kiyose Welfare Center for Handicapped　ii Graduate School of Information Science and Technology, The University of Tokyo,　iii Graduate School of Frontier Sciences, The University of Tokyo ,iv International Medical Center of Japan, Department of Otolaryngology, Tracheo-esophagology　v Prefectural University of Hiroshima, Faculty of Health and Welfare　vi Saitama Medical University, Faculty of Medicine

E-mail:　sakuraba@mtd.biglobe.ne.jp

Abstract　This study reports the use of a technique of automatic estimation of perceptual femininity in voice therapy for Male-to-Female transgender/transsexuals(MtF). The technique is based on speaker recognition techniques. Male models and female models are separately trained for vocal tract shape and pitch distribution. Using the four models, i.e., two genders and two parameters, perceptual femininity that is obtained through listening tests, is automatically estimated. The femininity estimation system is applied in actual voice therapy of MtF clients. Although it was accepted well to the clients, the estimation results sometimes are not fit to the femininity perceived by the therapist. This is mainly because the system captures only static acoustic properties of speech. This implies that the therapist can focus more on dynamic control of speech production in the therapy.

Keyword　Femininity，Automatic estimation，Speaker recognition，Gender identity disorder, Transsexual voice therapy

1. はじめに

著者ら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 transsexual voice therapy を行っており、今回の発表では知覚的女性度を推定する話者認識技術の臨床応用について検討する。

1.1. 性同一性障害

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, Tex Revision) [1] 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。

a)反対の性に対する強く持続的な同一感。
b)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。
c)その障害は、身体的に半陰陽を伴ったものではない。
d)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。

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性自認(gender identity)の問題であり、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）の問題とは区別されている。

日本では、1995 年より埼玉医大や岡山大が性同一性障害の診療を開始し、患者はホルモン療法や SRS（性別適合手術）などの医学的治療が受けられるようになった。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)が成立され、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、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

このように日本では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おり、2004 年に特例法が施行されてからは、生物学的性別をカムアウトすることなく、希望の性別で日常生活が送れるようになり、性同一性障害者の QOL は飛躍的に向上した。

1.2. MtF と声の問題

しかしながら、MtF の場合、容姿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めるものの、声に関してはホルモンや声帯手術による効果はほとんど望めない。櫻庭ら[2][3]は、話者に MtF が含まれていることを知らない第三者に、「Jack と豆の木」の朗読文を聞かせて、話者の性別を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。声帯手術施行者 4 名の聴取実験結果では、女声と判定された割合は 0～50%(ave.15%)しかなかった。それに比べて、80%以上を超えて女声と判定されたものは、生来声質が女声に近いものや狭・広義のボイストレーニング経験者であった。

2. Transsexual Voice Therapy

ボイストレーニングの方が声帯手術より、女声の獲得には有効な手段であることはわかったものの、日本における Transsexual Voice Therapy には未だ確立されたものがない。そこで、著者らは MtF のための Transsexual Voice Therapy 法を確立するために、音声データ採取や聴取実験を繰り返しながら、client のニーズに応じたセラピー法の検討を行っている[4]。

女性声の獲得のためには(1)高くなく低くない声の高さと、(2)女らしい話し方が必要である。女声の獲得のために、声の高さが重要であることはわかっているものの[5][6]、裏声やアニメ声など、単に男声の基本周波数(F0)を上げただけでは、櫻庭らの聴取実験では、女声判定率は 30%ほどしかなく、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た。櫻庭らの聴取実験では、180～230Hz あたりの声がもっとも女性らしく聞こえる高さであることがわかった。声の高さをあげる場合、男性声をそのまま裏声にするのではなく、喉を絞るようにして地声の一番高いところを引き伸ばすようにするといい。

女声獲得のために、声の高さ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)語尾をあげる (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、あまり技巧に走りすぎると、男性がわざとしゃべっているようにようにしか聞こえず、30%程度の女性度判定率でしかなくなる。

むしろ、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いる人の方が女性度判定率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る。

また、MtF の声に係わる医者やセラピストも「男が出す女声」という認識が頭から離れず、第三者の判定にかけると依然として男性と判断される声を、女性声として治療を終了させる場合もある[7]。

当事者や専門家以外で、私情や偏見を持たずに、当事者の声を判断してもらう聴取実験は声帯の外科的手術やセラピーの進捗度や成功度を測るためには、必須だと考えられる。しかしながら、偏見のない第三者に判定してもらうためには、常に新しい聴者を求める必要がある。また、判断には個人差がみられるので、統計にたえうるだけの聴者の人数を確保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を行うには、音声収録から、かなりの時間がかかってしまうことが多い。

このような聴取実験や評定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定可能なコンピュータによる知覚的性別判定を行うことは臨床的意義があると考えた。

3. 聴取実験による知覚的女声度の算出

第 4 節以降で使用する GID 話者の音声は基本的に全て聴取実験により、その“女らしさ”（知覚的女声度）が付与されている。まず、この聴取実験について述べる。

聴取実験は、本来の研究の目的を知らない聴者に、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させる方法で行った。聴取実験は数回に分けて行われ、1 回の聴取実験では約 20 名程度の音声を判定してもらった。手順は、まず最初に判定してもらう話者全部の「ジャックと豆の木」の朗読を聞かせた。次に「ジャックと豆の木」の冒頭 2 文を呈示し、(1)話者の年代（子供,10,20,30,40,50,60,70,80）、(2)話者の性別、(3)話し方の男らしさ、女らしさを判定させた。話し方の判定は段階評定尺度を用いた。段階評定尺度は 7 段階とし、話者の性別に関係なく、話し方が最も女らしいと感じたら 1 と評定し、最も男らしいと感じたら、7 に評定してもらった。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。聴者は各話者の各発話を 2 回判定したことになる。この聴取実験の結果、各話者について女性と判定されている割合が算出されるが、この割合を知覚的女性度と呼ぶ。

4. 女声度の工学的定義

音声情報処理技術の発展に伴い、言語情報以外のパラ・非言語情報の抽出技術について研究が行なわれている。例えば[9]では、話者認識技術を応用することで、話者の知覚的年齢の自動推定を試みている。

混合ガウス分布（GMM）を用いた話者モデリングは、ケプストラム系列へと変換された音声データ全体を一つの GMM でモデル化する。音声全体を対象とすることで音韻によるスペクトル変動はキャンセルされ、音声全体に対して静的に影響を及ぼす要因（即ち、話者性やマイクの特性など）のみがモデル化されることになる。話者 s のモデルを M_s とした場合、話者識別（観測量 o が予め用意された話者集団のどれかを判定する）は $P(o|M_s)$を算出することによって、話者照合（o が話者 s の音声であるか否かを判定する）は、$P(o|M_s)/P(o|M_{\neg})$を算出することで可能となる。本稿では基本的に話者照合の枠組みでの女声度の推定を考え、以下の式でこれを定義する。

$$F(o) = \log P(o \mid M_F) - \log P(o \mid M_M)$$

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。以下に示すようにパラメータとして声道形状に対応する MFCC、声の高さに対応する F_0 を用い、男女×2 パラメータの 4 つのモデルを構築した。

5. 女声度推定の実験

5.1. 実験に用いた音声

男性モデルおよび女声モデル作成には JNAS（新聞記事読み上げ音声コーパス）の音声から、男女それぞれ 114 名ずつ、各話者 30 発声を用いた。評価用の音声は GID 話者（19～78 歳）の音声 143 発声（うち、MtF111 名、FtM2 名の計 113 話者）である。本実験では、計算機により算出した女声度と第 3 節で算出した知覚的女声度との相関係数を求め、その性能の評価を行った。

5.2. 実験条件

本実験の分析条件を表 1 に示す。

表 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+Δ12 次元+ΔE 計 25 次元 logF0

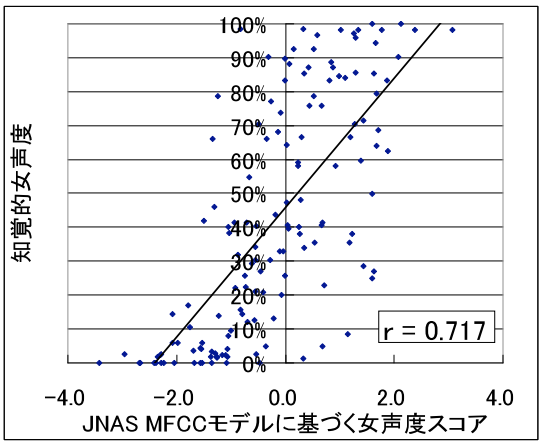

図 1 JNAS MFCC モデルに基づく女声度スコア 対 聴取結果

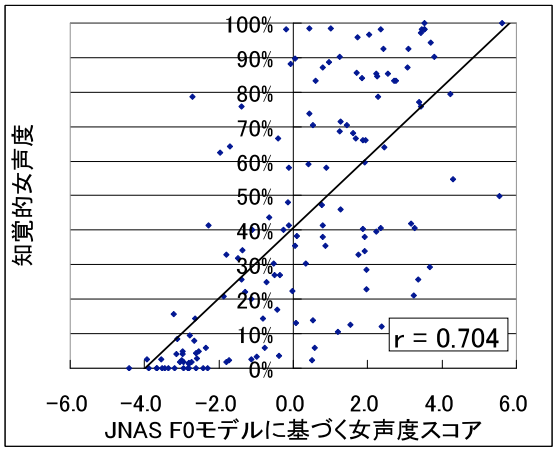

図 2 JNAS F_0 モデルに基づく女声度スコア 対 聴取結果

パラメータとしては、スペクトル包絡、すなわち声道形状の情報に相当する MFCC と、音源情報を表す $\log F_0$ の両方を用いた。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の除去を施した。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、実験に必要のない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。

5.3. モデル作成と女声度スコアの算出

続いて MFCC と $\log F_0$ の 2 つのパラメータそれぞれについて、女声モデルと男声モデルを 16 混合 GMM として作成した。$\log F_0$ 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らないことが期待できる。

以上のモデルに対し、GID 話者の音声を入力して女声度を算出した。このとき、GID 話者の音声の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 1、2 のようになった。

また、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。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。

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いかに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 が通常の女性のレンジに含まれている場合は女性と判定されてしまう。これは図 2 の右下にある音声に対応する。一方、MFCC をパラメータと使用した場合には、$\log F_0$ を使用したときと比べてばらつきが少なくなった。$\log F_0$ のモデルに比べ MFCC では女声度と女性判定率が乖離する音声が少なく、より安定した評価が可能であるといえる。また、前述の声の高さのみ高い音声についても、以前少数が図 1 の右下に存在するものの、多くは女声度が低いと適切に判定できた。

5.4. 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。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性モデルおよび女性モデルの対数尤度、あわせて 4 つの対数尤度のすべてを用いた。このときの散布図を図 3 に示す。このとき、相関係数は 0.799 となった。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。

しかし、この場合でも依然として外れサンプルが存在する。図 3 右下の音声は聴取実験では女性、計算機では男性と判定されたものであるが、第 5.3 節で触れたような、声が不自然に高いが、計算機がそれを女性らしいと高く評価してしまった音声が多く集まった。声の高さが不自然に高い音声の多くは、予測値もそれほど高くないものとして評価できたが、一部のものが依然評価が高いまま残ってしまったということである。図 3 左上の音声は聴取実験では女性、計算機では男性

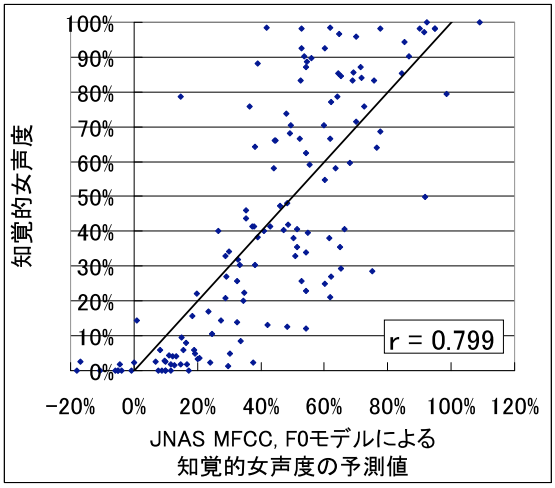

図 3 JNAS MFCC, F_0 モデルによる女声度スコアの予測値 対 聴取結果

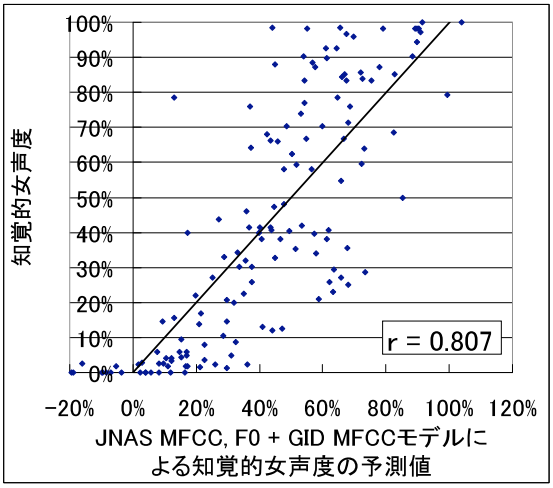

図 4 JNAS MFCC, F_0 + GID MFCC モデルによる女声度スコアの予測値 対 聴取結果

と判定されたものであるが、これらの音声は高齢の女性に聞こえるものが多数存在した。今回、モデル作成に用いたデータベースは JNAS の音声であったが、この話者は 20 代、30 代が多く、高齢の女性に聞こえるような音声は正しく評価できなかったと考えられる。

5.5. 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(0.717)に比べ、高い相関が得られた。

つぎに、第 5.4 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 5.4 節で用いた 4 つの対数尤度に、GID 話者モデルの MFCC に関する対数尤度 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。このときの予測値と聴取結果の散布図は図 4 のようになった。

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、わずかながらより高い相関が得られた。これは GID 話者による男声モデルの学習データに含まれるような、ただ声を高くした音声が低く評価されるようになり、また図左上に分布するような、生物学的女性の音声とは音響的に異なるものの、聴取では多くの人に女性と判定される音声がより高く評価されるようになったためということができる。

6. 知覚的女声度推定器の臨床応用

第 5 節で述べた女声度判定装置を 2006 年 2 月よりセラピーに随時導入して、client の反応や意見をみている。

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るので、声の低い MtF に高い声を出させる訓練をやる時には、数値が見えて有効であった。しかしながら、聴取実験においては、声の聴覚印象では声は高くなくても、話し方が女性らしい場合には 90%以上の聴者が女と判定している。知覚的女声度推定器では、そのような声に対しては女声度 6 割程度の判定しか出ておらず、今後、そのような声に対しても高い判定度が出るように改良する必要があることがわかった。また、声の高さは女声域であっても、話し方が女声らしくないものは、女声判定装置では高得点をマークする傾向にあるが、聴取実験では 50%を超えられない。故に、このような声に対しては女声度を現行より低く出すように改良する必要があることがわかった。

知覚的女声度推定器が女らしい話し方にも対応で

きるようにするためには、まず日本語における女らしい話し方を再考する必要がある。第 2 節でも述べたように、女性らしい話し方の特徴として(1)語尾を延ばす (2)語尾をあげる (3)抑揚に富む(4)軟起声を使用する (5)鼻音化するなどが言われているものの、極端にやりすぎると却って男声と判定されてしまい、声の高さ同様、適度な量で留める必要がある。その適度な量を算出することが今後の課題と言える。

現在、実際の臨床で知覚的女声度推定器を使用する際には、装置で判定できることと装置の限界を事前に client に説明し、納得してもらった上で使用することにしている。声の女声度が数値化して見えることで、訓練の目安や励みになっていることがわかった。

7. まとめ

MtF を対象とした音声がどの程度女性らしく聞こえるかを計算機により自動的に推定するため、話者認識技術に基づいた女声度を評価する手法を提案した。実験の結果から、人間が音声から話者の男女を判断するときには、声の高さだけではなく、フォルマントの情報も重要であることが示された。すなわち、より女性らしく聞こえる声を出すためには、ただ声を高くするだけでなく、声道形状を制御することも必要であるということである。また臨床応用のため、線形回帰分析を用いて聴取実験による女性判定率を予測した。予測値と実際の聴取結果との相関係数は 0.807 と良好な結果を得ることができた。

しかしながら、現在の知覚的女声度推定器では、話し方が女性らしいかどうかの判定はできないので、今後は女声らしい話し方を科学的に検討し、判定に加える必要があることがわかった。改良の余地は多いにあるものの、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。

***本研究の一部は JASE の助成金の支援を受けている。**

文　献

[1] 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

[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”, sp2002-187, 信学技法 vol.102 No.749, 2003

[3] 櫻庭京子他, “女声と聴取された性同一性障害者(MtF)の音声の音響分析”, 音講論（春）, 449-450,2003.

[4] 櫻庭京子他, “男性から女性に性別の移行を希望する性同一性障害者 MtF の Transsexual Voice Therapy”, 2005, 第 25 回性科学学会、口頭発表

[5] M.P.Gelfer and K.J.Schofield, “Comparison of acoustic and perceptual measures of voice in male-to-female transsexuals perceived as female vs. those perceived as male”, Journal of voice, 14, 22-33, 2000

[6] Wolfe,V.I., ‘Intonation and fundamental frequency in MtF TS’, J.Speech Hear Dis., 55, 43-50, 1990

[7] 櫻庭京子から共同研究者及び阿部輝夫医師、針間克己医師へのボイスセラピーに関する私信報告, 2006

[8] 丸山和孝他, “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”,音講論（春）, 3-P-20, 2006

[9] 峯松信明他, “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

[10] J. P. Campbell, “Speaker Recognition: A Tutorial,” Proc. IEEE, vol.85, No.9, pp.1437-1462, 1997