Vol.2013-MUS-99 No.47
2013/5/12

# 声から身体情報を求める

小林 真優子[1,a)]　西村 竜一[1,b)]　入野 俊夫[1,c)]　河原 英紀[1,d)]

**概要**：声を聴くと，何となくその人の体型が分かる．ここでは，母音だけを用いて相対的な声道長を推定する方法を提案する．この方法では，声道長以外の要因によるスペクトル形状変化の影響を軽減するために，スペクトル距離の計算に用いる帯域を制限し，スペクトルの大局的な平坦化と形状の過度な詳細の平滑化とを組合せている．6 歳から 56 歳までの 284 名の男女が発声した母音と身体情報からなるデータベースを用いることで，これらの処理に用いるパラメタを決定した．母音だけを用いた簡易な方法にも関わらず，以前報告した聴覚モデルを用いた方法を凌駕する精度での声道長推定が可能であることを確認した．また，このデータベースに付与された身体情報を母音だけから推定できることを示した．

**キーワード**：声道長正規化，音声，母音区間，スペクトル距離

# Voice tells your body information

Kobayashi Mayuko[1,a)]　Nisimura Ryuichi[1,b)]　Irino Toshio[1,c)]　Kawahara Hideki[1,d)]

***Abstract:*** When we hear a voice, we will see the person's body type somehow. In this article, we propose a method for estimating relative vocal tract length using only vowels. The proposed method consists of procedures to alleviate spectral deforming effects caused by other factors than the vocal tract length. They are selection of spectral region for calculating spectral distance, removal of global spectral shape, and smoothing of excessive details of spectrum. Parameter tuning of the proposed method was conducted by using a speech database with relevant physical data which consists of Japanese five vowels spoken by 284 male, female and adolescent talkers ranging from 6 to 56 years old. This simple vowel-based method found to provide better estimates than our previously proposed method. The proposed method also provides estimates of talkers' height and weight only from vowels using the relevant physical data stored in the database.

***Keywords:*** VTLN, Voice, vowels, spetral distance

## 1. はじめに

母国語の場合，日常の様々な環境における人間による音声の認識精度は，自動音声認識システムを大きく上回る [1], [2]．雑音や残響等の妨害の多い日常環境での音声コミュニケーションでは，それらの妨害を排除するとともに，様々に異なる話者の音声に迅速に適応することが必要となる．そのような人間の能力を支えている聴覚の機能として，まず，初期聴覚系における音を発する音源の形状と寸法の分離知覚 [3], [4] を挙げることが出来る．このように形状と寸法を分離して知覚することができるため，子供と大人のように体の寸法が大きく異なる場合でも，物理的性質が大きく異なるそれぞれが発声した母音（例えば「あ」）を、同じものとして知覚することができると考えるのである．さらに，未知の話者の音声への人間の迅速な適応能力を挙げることができる．百名の未知話者により孤立発声された単音節の知覚実験の結果によれば，5 個の単音節を聴くだけで，人間は未知の話者の音声に対する音声知覚機構の適応を完了させていることが示されている [5], [6]．このことは，未知話者への適応が音節に含まれている母音のみ

[1] 和歌山大学
Wakayama Uniersity, Wakayama 640–8510, Japan
[a)] s130043@center.wakayama-u.ac.jp
[b)] nisimura@sys.wakayama-u.ac.jp
[c)] irino@sys.wakayama-u.ac.jp
[d)] kawahara@sys.wakayama-u.ac.jp

を手掛かりとしていることを強く示唆する．

これらの知見に基づき，ここでは母音のみを用いた相対的声道長の推定法を提案する．提案する方法では，声道長比の推定に用いるスペクトルとして，著者らによる分析時刻に依存しないスペクトル包絡（TANDEM-STRAIGHTによるスペクトル包絡 [7], [8]）を用いている．こうして求められた異なった話者によるスペクトルを比較する際に，提案方法では，声道長以外の要因によるスペクトルの変形を回避するための 3 種類の処理を用いる．それらは，(1) スペクトル距離評価を行う周波数帯域の選択，(2) スペクトルの大局的形状の平坦化，(3) スペクトルの過度の詳細の平滑化である．これらの処理を組み合わせ，幅広い年齢層にわたる日本語母音のデータベース [9] を用いて処理パラメタを調整することにより，声道長比推定の標準誤差が 0.9%となることが示された．用いた資料が異なるために直接比較することはできないが，聴覚モデル [10] を用いた以前の報告 [11] を大きく上回る精度である．

## 2. 研究背景

声道長正規化は，自動音声認識システムなどに広く用いられている．しかし，スペクトル形状は，声道長だけでなく，個人差や発達による声道の形状変化 [12]，梨状窩による零 [13]，声帯音源波形によるいわゆる glottal formant によるピークや声門閉止区間の存在による周期的な零点および声門閉止の状況による高域でのスペクトル傾斜の変化 [14], [15]，高域での声道の 3 次元形状による多数の固有モード [16] のような他の要因に影響されて変形する．以前の報告 [11] では，周波数帯域選択と動的圧縮型ガンマチャープフィルタバンク (dCGCFB) に基づく聴覚モデル [10] を組み合わせることにより，これらの妨害要因の影響を軽減し，信頼性の高い声道長比推定が実現できることを示した．

しかし，この方法では比較対象となる異なった話者間の音声の時間軸整合を行う必要があり，利用できる状況が制限されるという問題があった．ここでは，母音のみを用いることでこの制限を回避するとともに，聴覚モデルを短時間フーリエ変換に基づくスペクトル表現に置き換え，周波数帯域選択に加えて，その他の声道長比推定における妨害要因を明示的に排除することで，簡易で高速な推定法を実現することを狙う．

### 2.1 身体情報付き母音データベース

ここで用いたデータベースには，標準語方言を話す 6〜56 歳までの男性，女性，子どもの音声が収録されている．具体的には，6〜56 歳までの男性話者 186 名と，6〜47 歳までの女性話者 199 名である．音声の収録には無指向性ミニチュアコンデンサマイクロフォン (DPA-4061) が用いられている．できるだけ自然な発話の資料とするために，「あれは X ぱんだ」というキャリア文に埋め込まれて発声された音節「は」「ひ」「ふ」「へ」「ほ」が切出されて収録されている．このデータベース中の 284 名 (男性 146 名，女性 138 名) には，音声だけではなく，身体情報（身長と体重）が併せて記録されている．

## 3. 母音を用いた声道長比推定法

ここでは，提案手法について説明する．この提案手法は大きく分けて 3 つの手順（母音テンプレートの作成，スペクトルの平滑化や平坦化，スペクトル比推定の距離尺度の最小化）により構成されている．下記に詳細を紹介する．

### 3.1 母音テンプレートの作成

まず，母音テンプレートを計算するために用いる安定した母音区間を，与えられた音声試料から選択する．こうして選択された区間内のフレーム毎に，分析時刻に依存しないスペクトル表現 $P_S(\omega,t)$ が，パワースペクトル $P(\omega,t)$ から次式で示される F0（基本周波数）適応型処理により求められる．

$$P_S(\omega,t) = \frac{1}{\omega_0}\int_{-\omega_0}^{\omega_0} h(\lambda)P(\omega-\lambda,t)d\lambda \ , \tag{1}$$

$$h(\omega) = \begin{cases} 1-\left|\frac{\omega}{\omega_0}\right|, & (|\omega|\le\omega_0) \\ 0, & (|\omega|>\omega_0) \end{cases} , \tag{2}$$

ここで $\omega_0 = 2\pi f_0$ は基本角周波数である．パワースペクトルを計算するための時間窓は，個々の高調波成分を分離するのに十分な長さに設定される．$k$ 番目の話者の母音テンプレート $G^{(k)}$ は各母音の平均対数パワースペクトル $L^{(v,k)}(\omega)$ の集合として定義される．ここで添字 $k$ は，話者を表し，添字 $v \in V = \{/\mathrm{a}/, /\mathrm{i}/, /\mathrm{u}/, /\mathrm{e}/, /\mathrm{o}/\}$ は，母音の種類 (音素) を表す．

$$G^{(k)} = \left\{ L^{\left(/\mathrm{a}/,k\right)}(\omega), L^{\left(/\mathrm{i}/,k\right)}(\omega), L^{\left(/\mathrm{u}/,k\right)}(\omega), L^{\left(/\mathrm{e}/,k\right)}(\omega), L^{\left(/\mathrm{o}/,k\right)}(\omega) \right\} , \tag{3}$$

$$L(\omega) = \frac{1}{\#(F)}\sum_{n\in F} 10\log_{10}\left(P_S(\omega,t(n))\right) \ , \tag{4}$$

ここで $F$ は特定の話者と特定の母音に対応するフレームの集合を表す．関数 $\#(F)$ は集合 $F$ の基数を表し，$t(n)$ はフレーム $n$ の時刻を表す．

### 3.2 スペクトルの平滑化と平坦化

距離計算に先立ち，声道長比較における不要なスペクトル変化の影響を排除するための処理を行った．大局的なスペクトル形状の平坦化と，過剰な細部の平滑化である．

大局的な平坦化の処理は，声道伝達特性のスペクトルの大局的な傾斜が本質的にはゼロであるという性質に基づいている．大局的な形状を等化されたスペクトル

$P_E^{(v,k)}(\omega)$ はテンプレートの構成要素であるスペクトル $P^{(v,k)}(\omega) = 10^{(L^{(v,k)}(\omega)/10)}$ から次式で求められる．

$$P_E^{(v,k)}(\omega) = \frac{P^{(v,k)}(\omega)\int_{-\omega_W}^{\omega_W} w_G(\lambda)d\lambda}{\int_{-\omega_W}^{\omega_W} w_G(\lambda)P^{(v,k)}(\omega-\lambda)d\lambda}\ , \qquad (5)$$

ここで $w_G(\omega)$ は大局的なスペクトル形状を求めるための平滑化関数である．関数の定義域の幅である $2\omega_W = 4\pi f_W$ は，フォルマント周波数の平均間隔よりも 2〜4 倍広い．現在の実装においては，平滑化関数として raised cosine 関数 $(1+\cos(\pi\omega/\omega_W))$ を用いている．

次の平滑化処理は，前に挙げた様々な要因により生ずる零点の影響やスペクトル形状の変形と，フォルマントに対応するピークの帯域幅の違いによる影響を軽減することを狙っている．正規化されたスペクトル $P_N^{(v,k)}(\omega)$ は，大局的形状を取り除いたスペクトル $P_E^{(v,k)}(\omega)$ から計算される．

$$P_E^{(v,k)}(\omega) = \frac{P^{(v,k)}(\omega)\int_{-\omega_W}^{\omega_W} w_G(\lambda)d\lambda}{\int_{-\omega_W}^{\omega_W} w_G(\lambda)P^{(v,k)}(\omega-\lambda)d\lambda}\ , \qquad (6)$$

ここで $w_N(\omega)$ はスペクトルの余分な詳細を取り除くための平滑化関数である．平滑化に用いる核関数の幅である $2\omega_N = 4\pi f_N$ は，通常のフォルマント帯域幅の数倍程度に設定する．現在の実装においては，平滑化関数として raised cosine 関数 $(1+\cos(\pi\omega/\omega_N))$ を用いている．これらの平滑化関数の定義域の広さを表すパラメタ ($\omega_W$ and $\omega_N$) は, 声道長比を推定するために用いられるスペクトル距離の値に影響を与える．

### 3.3 スペクトル距離の最小化

最後の処理は，テンプレート間のスペクトル距離計算に用いる周波数範囲の選択である．距離計算では，中央の周波数帯域だけが用いられ，高い周波数領域と低い周波数領域は排除される。それらの排除された周波数領域では，スペクトル形状の変化に対する声道長の影響よりもそれ以外の要因による影響の方が大きいためである．

$$d(k,n;a) = \left(\frac{1}{\#(V)\#(B)}\sum_{v\in V}\sum_{\omega_m\in B}\left|L_N^{(v,k)}(\omega_m) - L_N^{(v,n)}(a\omega_m) - \overline{L_N^{(v,k)}} + \overline{L_{N,a}^{(v,n)}}\right|^2\right)^{\frac{1}{2}}, \quad (7)$$

ここで
$$L_N^{(v,k)}(\omega) = 10\log_{10}\left(P_N^{(v,k)}(\omega)\right)\ ,$$
$$\overline{L_N^{(v,k)}} = \frac{1}{\#(B)}\sum_{\omega_m\in B} L_N^{(v,k)}(\omega_m)\ ,$$

ここで $B = \{\omega_L, \ldots, \omega_m, \ldots, \omega_H\}$ は，$\omega_L = 2\pi f_L$ から $\omega_H = 2\pi f_H$ の間に対数周波数軸上で等間隔に配置された離散周波数の集合を表す．現在の実装では，オクターブ毎に 24 個配置する間隔を用いている．最小スペクトル距離 $d_{\min}(k,n;a_{k,n})$ と声道長比の推定値 $r_{k,n} = l_k/l_n$ ($l_k$ と $l_n$ は $k$ 番目の話者と $n$ 番目の話者の声道長である) は，次式に示す最小化により求められる．

$$d_{\min}(k,n;a_{k,n}) = \underset{a}{\operatorname{argmin}}\, d(k,n;a)\ , \qquad (8)$$

ここで最小化を行うために操作されるパラメタ $a_{k,n}$ は，周波数軸の伸縮率を表す．周波数軸の伸縮は声道長の伸縮に反比例するため，スペクトルに基づく推定声道長比 $r_{k,n}$ は，$r_{k,n} = 1/a_{k,n}$ として求められる．距離評価周波数領域の境界周波数 $f_L$ と $f_H$ は，前述の平滑化関数の幅と同様に，提案法の精度に影響を与えるパラメタである．

## 4. 性能に影響するパラメタの調整

提案手法には，4 つのパラメータ $f_W$, $f_N$, $f_L$, $f_H$ が含まれている．声道長比の精度はこれらのパラメタに依存するため，適切な値に設定する必要がある．声道長の真の値を知ることは（実際的には）不可能であるため，ここでは，回帰分析により求めた推定値を用いることにより，これらの性能に影響するパラメタを調整することとした．以下の手順を用いてスペクトルに基づいて推定された声道長比を統合することにより，相対的声道長の（利用できるものとしては最良の）近似値が求められる．なお，この近似値を用いた評価法は、これまでの方法 [11] において用いていたものと同じである。

### 4.1 声道長比からの相対的声道長の推定

対数変換することで声道長比の計算は，線形化される．このことを利用すると，相対的な声道長を，以下の連立一次方程式の解から求めることができる．

声道長比を対数変換したものを全て並べたものを，ベクトル $\boldsymbol{r}$ とする．データベースに収録されている話者の声道長を（幾何平均が 1 になるように）正規化し対数変換したものを全て並べたものを，ベクトル $\boldsymbol{l}$ とする．このように定義すると，以下に示す接続行列 $\boldsymbol{H}$ を用いることにより，$\boldsymbol{r}$ を次式のようにモデル化することができる．なお，幾何平均を 1 とする正規化条件は，$\boldsymbol{H}$ の最終行として加えられている．

$$\boldsymbol{r} = \boldsymbol{H}\boldsymbol{l} + \boldsymbol{n}\ , \qquad (9)$$

ここで，$\boldsymbol{n}$ は観測誤差を表す．ベクトル $\boldsymbol{r}$, $\boldsymbol{l}$ および，接続行列 $\boldsymbol{H}$ は，具体的には次式により定義される．

$$\boldsymbol{r} = [\log(r_{1,2}), \log(r_{1,3}), \ldots, \log(r_{k,p}), \ldots, \log(r_{N,N-1}),\ 0]^T \qquad (10)$$

$$\boldsymbol{l} = [\log(l_1), \log(l_2), \ldots, \log(l_N)]^T \qquad (11)$$

$$\{\boldsymbol{H}\}_{m,n} = \begin{cases} 1 & (m=k) \\ -1 & (n=p) \\ 0 & (m\neq k)(n\neq l) \end{cases} \quad (12)$$

$p$ と $k$ は $\{\boldsymbol{r}\}_m = \log(r_{k,p})$ による。

$$\boldsymbol{H}|_{\text{last}} = [1,1,\ldots,1], (\text{正規化条件}), \quad (13)$$

ここで $N$ は話者の数を表し，$\{\boldsymbol{H}\}_{m,n}$ は $\boldsymbol{H}$ の $m$ 行 $n$ 列の要素を表す．また，$\boldsymbol{H}|_{\text{last}}$ は $\boldsymbol{H}$ の最終行を表す．最小自乗解 $\hat{\boldsymbol{l}}$ から，相対的な声道長の真の値の近似値 $\hat{l}_k$ が求められる．

$$\hat{\boldsymbol{l}} = (\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{R} \quad (14)$$

$$\hat{l}_k = \exp\left(\left\{\hat{\boldsymbol{l}}\right\}_k\right), \quad (15)$$

ここで $\left\{\hat{\boldsymbol{l}}\right\}_k$ はベクトル $\hat{\boldsymbol{l}}$ の $k$ 番目の要素を表す．

### 4.2 スペクトルに基づく推定の評価尺度

相対的な長さ $\hat{l}_k$ と $\hat{l}_p$ の真の値の近似値を用いて，比の真の値の近似値 $\hat{r}_{k,p} = \hat{l}_k/\hat{l}_p$ を定義する．この近似値とスペクトルに基づく声道長比の推定値の差の自乗平均値を声道長比の標準偏差により正規化したものとして，スペクトルに基づく推定の総合的な評価値 $\eta(f_W, f_N, f_L, f_H)$ を定義する．

$$\eta(f_W,f_N,f_L,f_H) = \left(\frac{\sum_{k\in S}\sum_{p\in(S-\{k\})}|r_{k,p}(f_W,f_N,f_L,f_H)-\hat{r}_{k,p}(f_W,f_N,f_L,f_H)|^2}{\sum_{k\in S}\sum_{p\in(S-\{k\})}\left|\hat{r}_{k,p}(f_W,f_N,f_L,f_H)-\overline{\hat{r}(f_W,f_N,f_L,f_H)}\right|^2}\right)^{\frac{1}{2}} \quad (16)$$

$$\overline{\hat{r}(f_W,f_N,f_L,f_H)} = \frac{1}{N(N-1)}\sum_{k\in S}\sum_{p\in(S-\{k\})} r_{k,p}(f_W,f_N,f_L,f_H) \quad (17)$$

ここでは式の見かけが複雑になるが，スペクトルに基づく推定値 $r_{k,p}$ と，回帰により求めた推定値 $\hat{r}_{k,p}$ が，性能に影響を与えるパラメタの関数であることを示すために，それらのパラメタの組 $(f_W, f_N, f_L, f_H)$ を式中に明示した．記号 $S$ は話者を表す添字の集合であり，$S-\{k\}$ は $k$ 番目の話者を除いた添字の集合である．

### 4.3 スペクトルに基づく推定法の調整

スペクトルに基づく推定でのスペクトル距離の最小化の処理を，Matlab の非線形最適化関数 `fminsearch` に含まれている simplex 法 [17] を用い，停止条件を $10^{-3}$ として実装した．母音データベースを用いた予備試験では，$f_W = 2000$ Hz, $f_N = 600$ Hz, $f_L = 400$ Hz, $f_H = 3500$ Hz としたときに最も良い結果が得られた．このパラメタの組み合わせを用いたときの総合的な評価値 $\eta$ の値は，0.12 であった．また，そのときのスペクトルに基づく声道長比の標準誤差は，0.0088 であった．この標準誤差は，評価に用いた資料が異なるが，以前に提案した方法 [11] の半分以下の大きさにあたる．

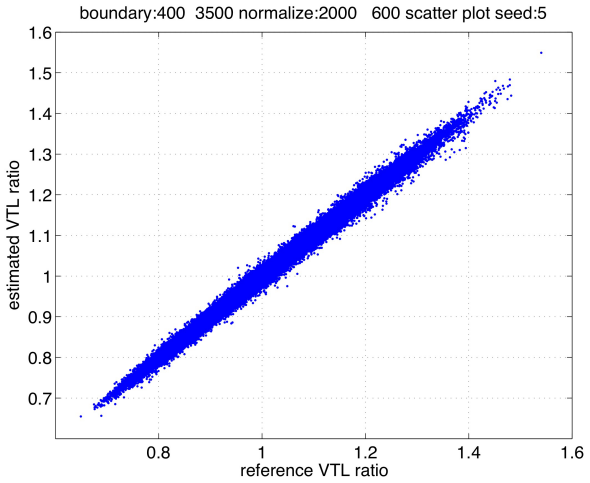


**図 1** Scatter plot of the regression-based VTL ratio (horizontal axis) and the spectrum-based VTL ratio (vertical axis).

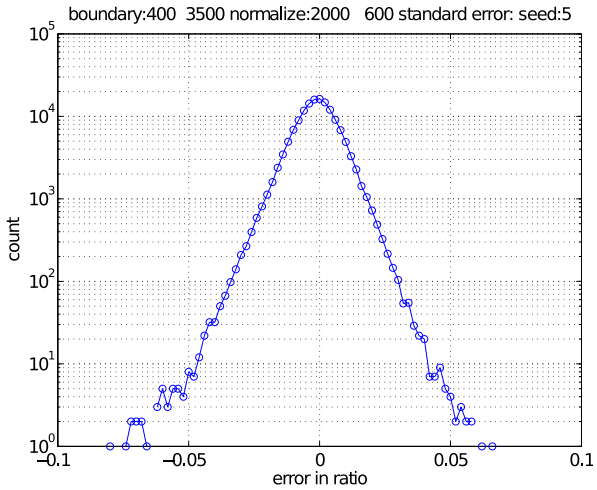


**図 2** Histogram of the spectrum-based VTL ratio estimation error.

図 1 に，回帰により求められた声道長比の推定値 $\hat{r}_{k,p}$ と，スペクトルに基づいて推定された声道長比 $r_{k,p}$ の散布図を示す。縦軸がスペクトルに基づく声道長比である。対角線上に点が集中しており，目立つ外れ値は無い．図 2 に，回帰により求められた値 $\hat{r}_{k,p}$ を正解とした場合の声道長比の推定誤差のヒストグラムを示す．話者の組み合わせの総数は 147840 である．この図から，スペクトルに基づいた声道長比の大部分が ±5%の誤差の範囲内にあることが分かる．

## 5. 母音からの身長・体重の推定

相対的な声道長の回帰分析による推定値 $\hat{l}_k$（以下、この章では声道長の推定値と略記する）は，正規化された実際の声道長の良い近似値となっていると考えることができる．ここでは，身体情報との関係を調べることとした．

まず，図 3 に，データベースに収録されている話者の年齢と身体情報の散布図を示す．上の図は身長，下の図は体

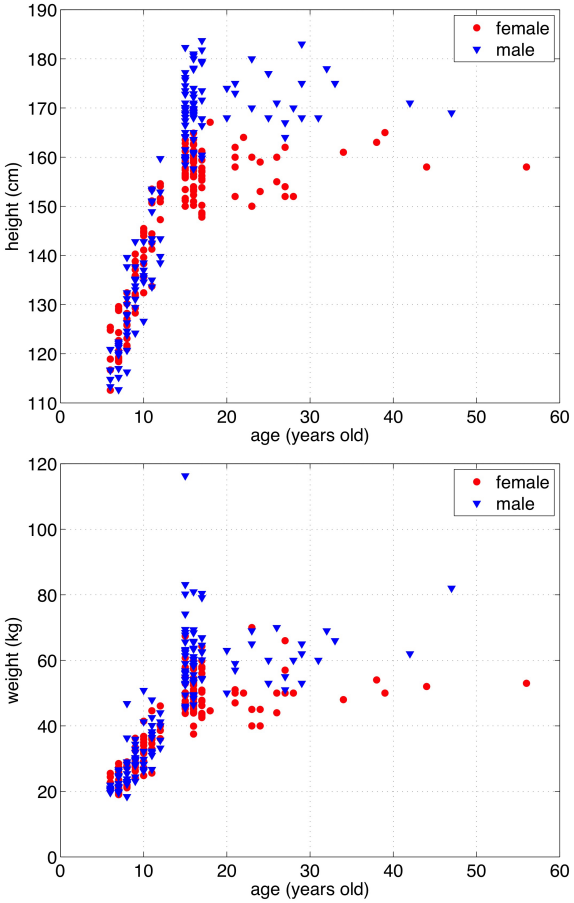


**図 3** Relation between speakers' age and their height (top plot) and relation between speakers' age and their weight (bottom plot).(from [18])

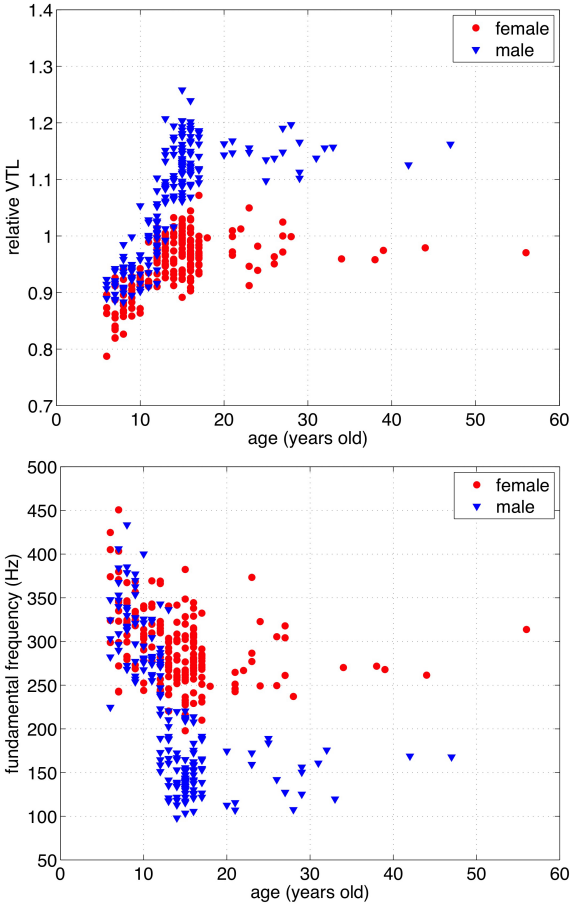


**図 4** Relation between speakers' age and their estimated relative vocal tract lengths (top plot) and relation between speakers' age and their average fundamental frequencies (bottom plot).(from [18])

**表 1** Summary of linear regression analysis for hight.(from [18])

| | height (cm) | | |
|---|---|---|---|
| | VTL | intercept | std. error |
| male | 178.905 | -34.646 | 7.614 |
| female | 208.78 | -48.64 | 8.949 |

**表 2** Summary of linear regression analysis for weight. (from [18])

| | weight (kg) | | |
|---|---|---|---|
| | VTL | intercept | std. error |
| male | 146.653 | -106.837 | 9.292 |
| female | 178.16 | -125.80 | 7.773 |

重の散布図である．12〜14 歳の話者については，音声データのみが収録されており，身体情報は提供されていない．これらの図では，話者の性別をマークの色と形で示している．赤丸が女性，青い三角が男性を表す．

図 4 に，年齢と音声の分析により求められる量の散布図を示す。上の図は，声道長の推定値，下の図は，基本周波数の散布図である．18 歳以下だけに注目すると，これらの量は年齢と単調な関係がある．図 3 の身体情報も，同じ範囲では年齢と単調な関係にある．これらは，この年齢との単調な関係を通じて，音声の分析により求められる量と身体情報を対応付けできる可能性があることを示唆している．なお，この章で示した音声の分析により求められる量は，スペクトル距離の最小化に simplex 法を用いる前の実装で求められたものである [18]．この場合の推定精度は若干低下するが，本質的な傾向に影響は無い．

図 5 に，推定された声道長と身体情報の散布図を示す。上の図は身長，下の図は体重との散布図である．これらを用いて，それぞれの性別毎に，推定された声道長を説明変数として身体情報を目的変数とした回帰分析を行った．表 1 と表 2 に結果を示す．表 1 は身長，表 2 は体重を目的変数とした分析結果である．全ての係数と切片は有意であり，母音のみから身体情報を推定できることを示している．図 5 には，こうして求められた回帰直線を記入した．なお，独立変数として基本周波数を加えた重回帰分析を行ったところ，期待に反して身体情報の推定精度の向上は認められなかった．この結果は，推定された声道長と基本周波数が高い相関を有しているためであると考えられる．

## 6. おわりに

母音のみに基づいて声道長比と身体情報を推定する新し

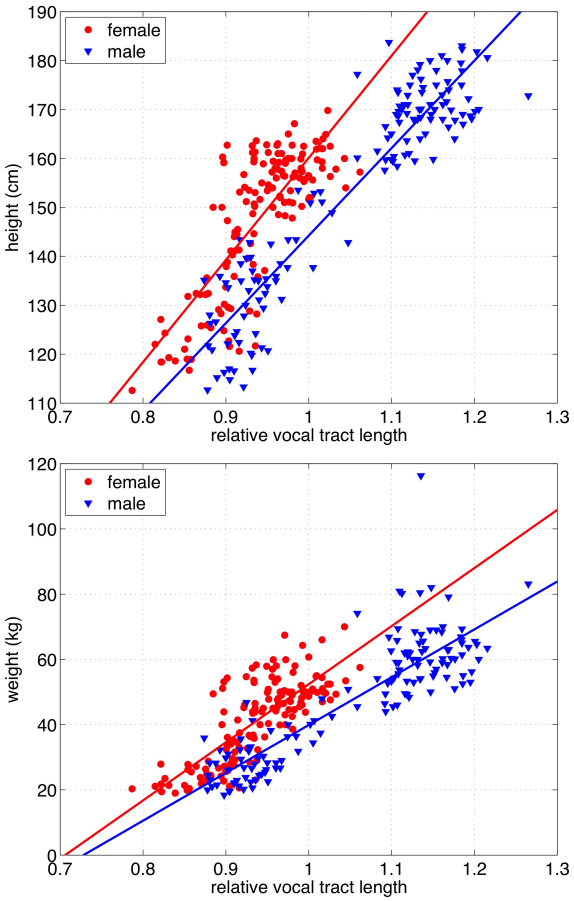


**図 5** Relation between the estimated relative vocal tract length and the speakers' height (top plot) and relation between the estimated relative vocal tract length and the speakers' weight (bottom plot). Lines in the plots represent the linear regression results.(from [18])

い方法を提案した．提案法は，短時間フーリエ変換に基づく簡易な方法であるにも関わらず，声道長比を 0.9%の標準誤差で推定することができる．また，身長や体重などの身体情報が付与された広い年齢層にわたる母音データベースにこの方法を用いることにより，同様に母音から身体情報を推定できることが示された．提案法は，FFT と線形補間という計算量の少ない処理を用いて実装されている．この高い精度と効率の良い実装は，母音に基づく音声変換，音声認識，話者認証など様々な応用に用いる際に有用な提案法の特徴である．

**謝辞** 本研究の一部は，科学研究費基盤 (B)24300073 および萌芽 24650085 による．

## 参考文献

[1] Saon, G. and Chien, J.-T.: Large-Vocabulary Continuous Speech Recognition Systems: A Look at Some Recent Advances, *Signal Processing Magazine, IEEE*, Vol. 29, No. 6, pp. 18 –33 (online), (2012).

[2] Stern, R. and Morgan, N.: Hearing Is Believing: Biologically Inspired Methods for Robust Automatic Speech Recognition, *Signal Processing Magazine, IEEE*, Vol. 29, No. 6, pp. 34 –43 (online), (2012).

[3] Irino, T. and Patterson, R. D.: Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The stabilised wavelet-Mellin transform, *Speech Communication*, Vol. 36, No. 3–4, pp. 181 – 203 (2002).

[4] Smith, D. R. R., Patterson, R. D., Turner, R., Kawahara, H. and Irino, T.: The processing and perception of size information in speech sounds, *The Journal of the Acoustical Society of America*, Vol. 117, No. 1, pp. 305–318 (2005).

[5] 加藤和美，筧　一彦：音声知覚における話者への適応性の検討，日本音響学会誌，Vol. 44, No. 3, pp. 180–186 (1988).

[6] Kakehi, K.: Adaptability to differences between talkers in Japanese monosyllabic perception, *Speech perception, production and linguistic structure* (Tohkura, Y., Vatikiotis-Bateson, E. and Sagisaka, Y., eds.), IOS Press, pp. 135–142 (1992).

[7] 森勢将雅，高橋　徹，河原英紀，入野俊夫：窓関数による分析時刻の影響を受けにくい周期信号のパワースペクトル推定法，電子情報通信学会論文誌 D，Vol. J 90-D, No. 12, pp. 3265–3267 (2007).

[8] Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T. and Banno, H.: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0 and aperiodicity estimation, *Proc. ICASSP2008*, pp. 3933–3936 (2008).

[9] 大山　玄，出口利定，粕谷英樹：幅広い年齢層にわたる日本語母音のデータベースの構築，日本音響学会春季研究発表会講演論文集，pp. 2–P–15(a) (2011).

[10] Irino, T. and Patterson, R.: A Dynamic Compressive Gammachirp Auditory Filterbank, *Audio, Speech, and Language Processing, IEEE Transactions on*, Vol. 14, No. 6, pp. 2222 –2232 (2006).

[11] Okamoto, E., Irino, T., Nisimura, R. and Kawahara, H.: Evaluation of voice morphing using vocal tract length normalization based on auditory filterbank, *J. Signal Processing*, Vol. 15, No. 4, pp. 283–286 (2011).

[12] Fitch, W. T. and Giedd, J.: Morphology and development of the human vocal tract: A study using magnetic resonance imaging, *J. Acoust. Soc. Am.*, Vol. 106, No. 3, pp. 1511–1522 (1999).

[13] Dang, J. and Honda, K.: Acoustic characteristics of the piriform fossa in models and humans, *J. Acoust. Soc. Am.*, Vol. 101, No. 1, pp. 456–465 (1997).

[14] Childers, D. G. and Ahn, C.: Modeling the glottal volume-velocity waveform for three voice types, *J. Acoust. Soc. Am.*, Vol. 97, No. 1, pp. 505–519 (1995).

[15] Fant, G. and Liljencrants, J.: A four-parameter model of glottal flow, *STL-QPSR*, Vol. 26, No. 4, pp. 1–13 (1985).

[16] Ternström, S. O.: Hi-Fi voice: observations on the distribution of energy in the singing voice spectrum above 5 kHz, *Proc. Acoustics'08 Paris*, pp. 3171–3176 (2008).

[17] Lagarias, J. C., Reeds, J. A., Wright, M. H. and Wright, P. E.: Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions, *SIAM Journal of Optimization*, Vol. 9, No. 1, pp. 112–147 (1998).

[18] Kobayashi, M., Nisimura, R., Irino, T. and Kawahara, H.: Estimated relative vocal tract lengths from vowel spectra based on fundamental frequency adaptive analyses and their relations to relevant physical data of speakers, *Proc. ICA/ASA*, International Congress on Acoustics (2013). (Accepted for publication).