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.

声道の局所的伸縮による話者正規化

北村　達也† 　竹本　浩典†† 　足立　整治†††

† 甲南大学知能情報学部
658-8501 兵庫県神戸市東灘区岡本 8-9-1
†† 情報通信研究機構ユニバーサルメディア研究センター
619-0288　京都府相楽郡精華町光台 2-2-2
††† Fraunhofer Institute for Building Physics
Nobelstrasse 12, 70569 Stuttgart, Germany
E-mail: †t-kitamu@konan-u.ac.jp, ††takemoto@nict.go.jp, †††seiji.adachi@ibp.fraunhofer.de

あらまし　Vocal tract length sensitivity function を用いて声道伝達関数の話者正規化を試みた．Vocal tract length sensitivity function とは，声道の長さ方向の局所的伸縮がフォルマント周波数に与える影響を求める関数である．これを用いて成人男性 6 名の日本語 5 母音の第 1 から第 4 フォルマント周波数が目標話者のものと一致するよう声道断面積関数を変形した．そして，元の声道断面積関数から変形後への写像として声道ワーピング関数を求めた．その結果，(1) 得られた声道ワーピング関数の形状は線形ではない，(2) 変形後の声道断面積関数の声道長は目標話者のものと異なる, (3) 各母音の声道ワーピング関数は話者内でも形状が異なる，ということが明らかになった．

キーワード　声道伝達関数, 母音, 話者間分散, Vocal tract length sensitivity function, 声道ワーピング関数

Speaker normalization by local expansion and contraction of the vocal tract

Tatsuya KITAMURA†, Hironori TAKEMOTO††, and Seiji ADACHI†††

† Faculty of Intelligence and Informatics, Konan University
8-9-1 Okamoto, Higashinada, Kobe, Hyogo 658-8501, Japan
†† NICT Universal Media Research Center
2-2-2 Hikaridai, Keihanna Science City, Kyoto 619-0288, Japan
††† Fraunhofer Institute for Building Physics
Nobelstrasse 12, 70569 Stuttgart, Germany
E-mail: †t-kitamu@konan-u.ac.jp, ††takemoto@nict.go.jp, †††seiji.adachi@ibp.fraunhofer.de

Abstract　Vocal tract warping functions for normalizing inter-speaker differences of vocal tract transfer functions were calculated based on the vocal tract length sensitivity function. The length sensitivity function is an equation for finding a change in formant frequency due to longitudinal perturbation of the vocal tract Vocal tract area functions for the five Japanese vowels of six male speakers were tuned for their first four formant frequencies to be close to those of a target speaker. The vocal tract warping functions were obtained as relationship between the original and deformed area functions. The results indicate that (1) the warping functions are not linear, (2) the vocal tract length of the deformed area functions are different from that of the target speaker, and (3) the shape of the warping functions of the five vowels are not constant for each speaker.

Key words　Vocal tract transfer function, Vowels, Inter-speaker differences, Vocal tract length sensitivity function, Vocal tract warping function

1. はじめに

音声における話者間分散は音声認識の精度向上を妨げる要因の 1 つであり，これを克服するために様々な研究が行われている．話者間分散は発話器官の生得的な個人差や後天的に獲得する話し方や方言などに起因する．中でも声道長の個人差は音声認識の特徴パラメータとして重要な音声スペクトルの話者間分散を生み出す要因の 1 つである．そのため，話者間の声道長を正規化する数々の手法が検討されてきた ([1] [2] [3] [4] など)．これらの手法では，周波数領域もしくはケプストラムの領域でスペクトルの周波数方向の伸縮を行うことにより声道長を正規化する．これは，現段階では音声から精度良く声道長を逆推定するのが困難なためである．一方，MRI(磁気共鳴画像法) を用いて発話時の声道形状を高精度に計測できるようになり [5]，声道の実測データを用いて話者正規化を試みる素地が整った．そこで，本研究では，MRI 観測により実測された声道形状を変形して話者正規化を行い，変形前後の写像関数 (声道ワーピング関数) を求めることを試みる．得られる声道ワーピング関数は，今後の話者正規化法の研究に何らかのヒントを提供できる可能性がある．

実測データにもとづく声道長正規化に関しては，Yang and Kasuya [6] は，成人男性，成人女性，子供の声道断面積関数を対象にした研究を行っている．彼らは，声道長を一様に伸縮させる一様正規化と，声道を喉頭腔，咽頭腔，口腔に分割し，それぞれの長さを正規化する非一様正規化の 2 つを試みている．なお，いずれの正規化法でも声道断面積の最大値が正規化されている．彼らは，この 2 つの手法で正規化した声道断面積関数の第 1，第 2 フォルマント周波数を比較し，主に声道全体の長さが声道の正規化に寄与すると報告した．また，北村ら [7] は成人男性話者 8 名の母音/i/と/e/を対象にして声道断面積関数を分析した．そして，声道長を一様に正規化しても声道伝達関数の話者間分散を吸収しきれないことを示している．

最近，Adachi ら [8] は vocal tract area sensitivity function と vocal tract length sensitivity function を利用して，所望のフォルマント周波数が得られるように声道を変形する手法を提案した．前者は断面積の変化がフォルマント周波数に及ぼす影響，後者は長さ方向の変化がフォルマント周波数に及ぼす影響を求めるための関数である．これらの関数を求めることによって所望のフォルマント周波数を得るには声道断面積関数のどの部分をどのように変形すべきか知ることができる．彼らは，この手法を用いて男女間の話者変換も可能なことを示した．

本研究では，Adachi らの手法を活用し，vocal tract length sensitivity function を用いて成人男性を対象とした話者正規化を試みる．そして，元の声道断面積関数から変形後の声道断面積関数への写像として声道ワーピング関数を求める．この関数の形状から，新たな話者正規化法の開発に寄与する知見を得ることを目標とする．近年では，音声合成における声質変換にも声道長正規化の技術が利用されているため，この分野に対しても何らかの寄与ができる可能性がある．

2. データ

2.1 MRI データ

成人日本人男性 7 名 (AN, HT, KH, SA, SH, TI, YT) が日本語 5 母音 (/a/, /i/, /u/, /e/, /o/) 発話中の MRI データを測定した．このうちの 2 名は文献 [7] の被験者である．使用した MRI 装置は ATR 脳活動イメージングセンタに設置された Shimadzu-Marconi ECLIPSE 1.5T Power Drive 250 である．MRI 装置内に仰臥位にて横になった被験者の頭頸部をヘッドネックコイルで撮像した．撮像条件を表 1 に示す．モーション・アーティファクトによるぶれが見られる画像は，以降の分析から除外した．

表 1　MRI の撮像条件
Table 1　Conditions of MRI measurement.

撮像シーケンス	Fast spin echo 法
Echo time (TE)	11 ms
Repetition time (TR)	3,000 ms
Flip angle (FA)	90°
撮像領域	256×256 mm (512×512 pixels)
スライス	矢状方向，厚み 2.0 mm，間隔なし
スライス枚数	41 または 51
加算回数	1 回

2.2 声道断面積関数

MRI では歯列は空気と同じく低輝度にしか造影できないため，MRI データにおいて歯列と空気の境界を決定するのが困難である．そこで，まず Takemoto ら [9] の手法を用いて MRI データに上顎と下顎のボリュームデータを補填した．

その後，Takemoto ら [10] の方法を用い，声道中心線に沿って 2.5 mm 間隔で声道の断面積を計測し，声道断面積関数を得た．両側の梨状窩は除外した．本研究では，声道を図 1 に示すような円錐台の連続体として近似する．

得られた声道断面積関数及び声道長をそれぞれ図 2，表 2 に示す．図 2 は 7 名の声道断面積関数は単純な相似形ではないことを示しており，表 2 は 7 名の声道長は最短 16.25 cm から最長 19.50 cm まで，最大 3 cm 以上の開きがあることを示している．

図 1　円錐台の連続体で表した声道断面積関数
Fig. 1　Schematic vocal tract area function represented by a succession of truncated cones.

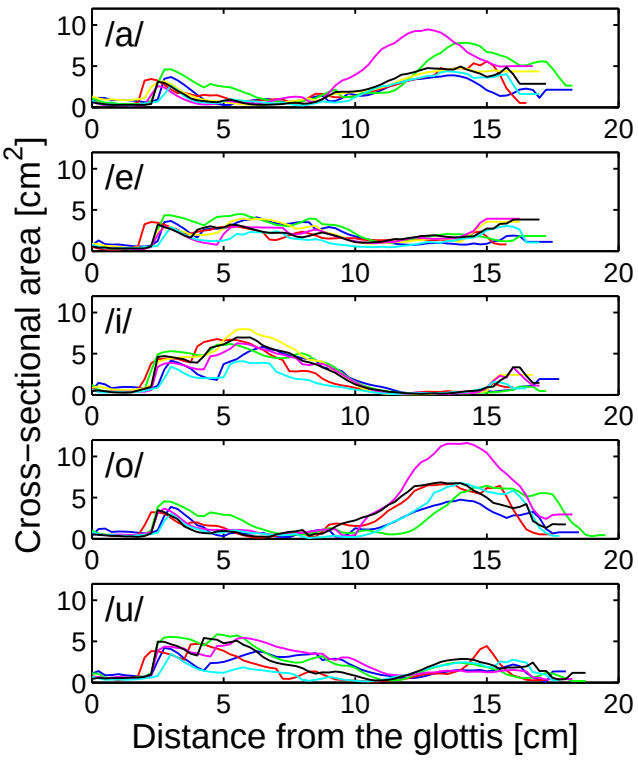

図 2　被験者 7 名の日本語 5 母音の声道断面積関数

Fig. 2　Vocal tract area functions of the five Japanese vowels from seven subjects.

表 2　被験者 7 名の日本語 5 母音の声道長 [cm]

Table 2　Vocal tract length of the five Japanese vowels from seven subjects in cm.

	Vowel				
Subject	/a/	/e/	/i/	/o/	/u/
AN	18.25	17.50	17.75	18.50	18.00
HT	16.50	15.75	16.00	17.50	17.25
KH	18.25	17.25	17.25	19.50	18.75
SA	17.00	16.25	16.75		
SH	16.75	16.25	17.00	18.25	18.25
TI	17.00	17.00	16.50	17.75	17.75
YT	17.25	17.00	17.00	18.00	18.75
Mean	17.29	16.71	16.89	18.25	18.13

音声認識技術においては，成人男性のみを対象にする場合でさえ，上記のような話者間分散のある声道から生成される音声に対応する必要がある．同様に，声質変換においては，上記のような話者間分散を表現できるようにする必要がある．

3. 方　　法

3.1 声道伝達関数の計算法

声道断面積関数の伝達関数は，等価回路モデル [11] により 5 kHz まで計算した．その際，声門開口面積は 0 と仮定し，口唇における放射特性 Z_R は，Caussé ら [12] により提案された以下の式により定めた．

$$\frac{Z_R}{\rho c} = \frac{z^2}{4} + 0.0127z^4 + 0.082z^4 \ln z - 0.023z^6 + j(0.6133z - 0.036z^3 + 0.034z^3 \ln z - 0.0187z^5) \quad (1)$$

$$z = kr \quad (2)$$

ここで，ρ は空気の密度，c は音速，k は波数，r は開口端の半径である．ρ は 1.15 kg/m^3，c は 349.3 m/sec を用いた．なお，上式は $kr < 1.5$ の条件下で有効である．本研究のデータでは開口端の半径の最大値が 11 mm であるため (被験者 YT の母音/e/)，5 kHz 以下において上式は有効である．

本研究で用いた等価回路モデルでは，上記の損失の他，熱伝導による損失，粘性摩擦による損失，声道壁の振動による損失を考慮してある．

3.2 Vocal tract length sensitivity function

Vocal tract length sensitivity function は，Adachi ら [8] により導出された，声道の長さ方向の変化によるフォルマント周波数への影響を求めるための関数である．

声道内の平面波伝搬を仮定することによって，声道を 1 次元の断面積関数 $A(x)$ で表すことができる．ここで x は声門からの距離である．声道内の平面波は，音圧 $p(x,t)$ と体積速度 $U(x,t)$ で記述される．声道内の気流は声道壁を通過することはないため，声道の第 n モードにおいて声道壁を押し広げる圧力は以下の式で与えられる．

$$P^{(n)}(x) = \mathrm{PE}^{(n)}(x) - \mathrm{KE}^{(n)}(x) \quad (3)$$

ここで，$\mathrm{PE}^{(n)}(x)$ は時間平均した位置エネルギー密度 (potential energy density) であり，$\mathrm{KE}^{(n)}(x)$ は時間平均した運動エネルギー密度 (kinetic energy density) である．これらのエネルギー密度は以下のように定義される．

$$\mathrm{PE}^{(n)}(x,t) = \frac{1}{2}\frac{1}{\rho c^2}p_n^2(x,t) \quad (4)$$

$$\mathrm{KE}^{(n)}(x,t) = \frac{1}{2}\rho\left(\frac{U_n(x,t)}{A(x)}\right)^2 \quad (5)$$

ここで，$p_n(x,t)$ と $U_n(x,t)$ はそれぞれ第 n モードが生じているときの音圧と体積速度である．第 n モードにおける全エネルギー E_n は，これらのエネルギー密度から求めることができる．

$$E_n = \int_0^L \left\{\mathrm{PE}^{(n)}(x) + \mathrm{KE}^{(n)}(x)\right\} A(x)dx \quad (6)$$

次に，図 3 に示すような声道の長さ方向の変形について考える [8]．この変形は，位置 x における断面積を長さ方向に $\delta x(x)$ だけ変位させるものである．ただし，$\delta x(0) = 0$ である．位置 x における局所的な伸縮の比率は

$$\Delta x \equiv \frac{\delta x(x)}{dx} \quad (7)$$

と表される．この変形により位置 x は $x' = x + \delta x(x)$ に移動し，変形後の声道断面積関数は

$$A(x') \equiv A(x) \quad (8)$$

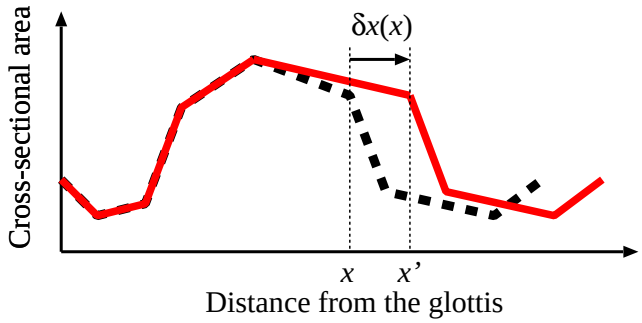

図 3　長さ方向の声道変形

Fig. 3　Longitudinal perturbation of vocal tract.

と定義される．このとき，第 n モード (第 n フォルマント周波数) における vocal tract length sensitivity function $S^{(n)}(x)$ は以下の式で定義される．

$$S^{(n)}(x) = -\frac{\left\{\mathrm{PE}^{(n)}(x) + \mathrm{KE}^{(n)}(x)\right\} A(x)}{E_n} \tag{9}$$

声道断面積関数を図 1 のような区分線形関数として表した場合，ノード (x_s, A_s) の集合体として離散的に表すことができる．ここで，s はノードの番号 $(s = 0, \ldots, N_s)$，x_s は声門からの距離，A_s は断面積である．なお，N_s は声道断面積関数の区間数である．このとき，離散形式の vocal tract length sensitivity function は以下の式で定義される．

$$\mathrm{PKE}_s^{(n)} = \mathrm{PE}_s^{(n)} + \mathrm{KE}_s^{(n)} \tag{10}$$

$$S_s^{(n)} = -\frac{\Delta x_s}{2E_n}\left(\mathrm{PKE}_s^{(n)} A_s + \mathrm{PKE}_{s-1}^{(n)} A_{s-1}\right) \tag{11}$$

3.3　Vocal tract length sensitivity function を利用した声道変形

Adachi ら [8] は，上記の vocal tract length sensitivity function に加え，従来からある vocal tract area sensitivity function を用いて，所望のフォルマント周波数を持つように声道形状を変形する手法も提案している(注1)．彼らの手法では，声道断面積関数の局所的な断面積の拡大/縮小及び区間長の伸張/短縮によりフォルマント周波数を調整している．

一方，本研究では，vocal tract length sensitivity function のみを用いて，すなわち声道断面積関数の局所的な伸張と短縮のみによってフォルマント周波数を調整する．将来的に，本格的な話者正規化を考える上では声道変形の自由度を制限しておいた方がよく，かつ従来の声道長正規化手法ともなじみがよいと判断したからである．

第 n フォルマント周波数の目標値を T_n，声道断面積関数 (x_s, A_s) から求められる第 n フォルマント周波数を f_n とする．これら 2 つのフォルマント周波数の差異を f_n で正規化した値は以下のようになる．

$$z_n = \frac{T_n - f_n}{f_n} \tag{12}$$

(注1)：Story [13] は 2006 年に vocal tract area sensitivity function を用いて，所望のフォルマント周波数を持つように声道形状を変形する手法を提案している．

表 3　被験者 7 名の日本語 5 母音の声道断面積関数から求めた第 1 から第 4 フォルマント周波数 (F1，F2，F3，F4) の平均値及び標準偏差 (SD) [Hz]

Table 3　Mean and standard deviation (SD) of the first, second, third, and fourth formant frequencies (F1, F2, F3, and F4) of vocal tract transfer functions of the five Japanese vowels from seven subjects in Hz.

	Vowel				
	/a/	/e/	/i/	/o/	/u/
F1 (Mean)	608	494	264	487	344
F1 (SD)	73	53	20	68	35
F2 (Mean)	1,127	1,870	2,374	788	1,189
F2 (SD)	64	109	179	72	204
F3 (Mean)	2,661	2,584	3,063	2,595	2,412
F3 (SD)	143	159	207	168	209
F4 (Mean)	3,501	3,454	3,753	3,473	3,517
F4 (SD)	221	179	388	141	140

そして，以下の規則にもとづき声道の微小変形を繰り返す．

$$x_s^{new} = x_{s-1}^{new} + \Delta x_s \left(1 + \beta \sum_{n=1}^{N_f} z_n S_s^n\right)$$
$$\text{for } s = 1, \ldots, N_s \text{ with } x_0^{new} = x_0, \tag{13}$$

ここで，N_f は調整するフォルマント周波数の数で，β は変形の度合いを調整する係数である．本研究では，N_f を 4，β を 2.0 と設定した．

Adachi ら [8] と同様，声道下部に位置する喉頭腔が極端に長くなることを避けるため，喉頭腔の区間では以下の制約を追加した．

$$x_s^{new} = \min\left\{x_{s-1}^{new} + \Delta x_{max},\ x_s^{new}\right\} \tag{14}$$

本研究では，Δx_{max} を 6.5 mm に設定した．

本研究では，式 (13) 及び (14) にもとづく変形を z_n $(n = 1, \ldots, 4)$ が 0.01 以下(注2)もしくは繰り返しが 1,000 回に達するまで繰り返した．

3.4　声道ワーピング関数の計算

声道ワーピング関数は x_s から x_s^{new} への写像関数である．本研究では，声道長が被験者間の平均値に近い被験者 AN を目標話者と設定した．そして，AN 以外の 6 名に関して第 1 から第 4 フォルマント周波数が AN のものに近づくよう声道断面積関数を変形した．

4.　結果と考察

被験者 7 名の 5 母音の声道伝達関数を図 4 に示す．これらの声道伝達関数からピークピッキングにより求めた第 1 から第 4 フォルマント周波数を表 3 に示す．声道伝達関数の個人差が周波数方向の一様なシフトでは表せないことがわかる．

声道ワーピング関数を図 5 に示す．横軸，縦軸ともに声門からの距離を表している．その始端 (声門) から約 2.5 cm までが

(注2)：Adachi ら [8] は，z_n のノルムが 0.01 以下になるまでというより厳しい条件を設定している．

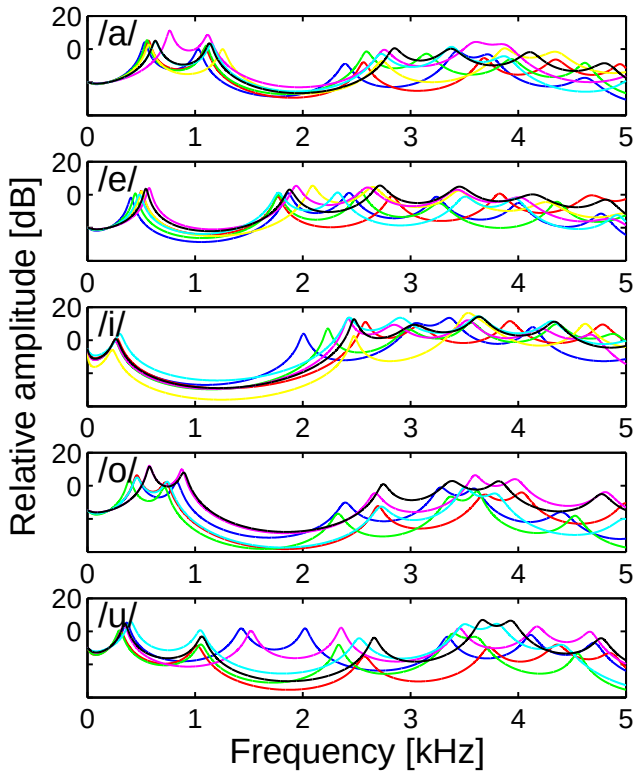

図 4　被験者 7 名の日本語 5 母音の声道伝達関数

Fig. 4　Vocal tract transfer functions of the five Japanese vowels from seven subjects.

喉頭腔，約 2.5 cm から約 10 cm までが咽頭腔，そして約 10 cm から終端 (口唇端) までが口腔に対応する．また，誤差の評価のため，z_n のノルム，つまり z_n の二乗和の平方根を表 4 に示す．いくつかの声道断面積関数に関して，1,000 回の変形後も z_n が閾値 0.01 を下回らないものがあった．具体的には，被験者 HT，KH，SA，SH，YT の母音/e/と被験者 SH の母音/u/の声道断面積関数において，z_n のいずれかもしくは複数が閾値を越えた．

図 5 の結果から，声道ワーピング関数が線形ではないことがわかる．これは，図 4 において声道伝達関数の個人差が周波数方向の一様なシフトでは表せないことからも予想されたことである．また，この結果は声道長の一様正規化では同性内の話者間分散を吸収できないことを示した北村ら [7] の結果とも一致する．

上記の結果と密接に関連するが，変形後の声道長が目標話者 AN の声道長に一致するわけではないこともわかる．この結果は，成人のフォルマント周波数の同性内の個人差を生み出す主要因が声道長ではなく声道断面積関数の凹凸パターンなどの個人差であることを示唆している．

また，各被験者の声道ワーピング関数は母音ごとに異なっており，声道長正規化のパラメータを話者内で統一できないことを示唆している．

5. おわりに

本研究では，声道の長さ方向の局所的伸縮によるフォルマント周波数の変化を求める関数である vocal tract length sensitivity function [8] を利用して，声道伝達関数の話者正規化を試みた．成人男性 6 名の声道伝達関数のフォルマント周波数が目標話者のものに近づくように声道断面積関数に局所的な伸縮を施した．そして，元の声道断面積関数から変形後への写像関数として声道ワーピング関数を得た．

変形後の声道断面積関数の声道長は目標話者のものと異なることから，(少なくとも) 同性内の話者正規化においては声道長の変更だけでは話者間分散を吸収できないことが示唆された．恐らく，従来の声道長正規化には声道長以外の個人性も正規化する効果があるものと考えられる．また，この結果は，音声合成における話者変換においても，単に声道長を伸縮するだけでは話者の違いを十分に生成できないことを示唆している．

また，得られた声道ワーピング関数の形状は母音ごとに異なり，この手法では声道長正規化のパラメータを話者内で一意に決められないことを示している．これは応用を考える上で大きな障壁となるので今後の検討を要する．

表 4　z_n のノルム．z_n はフォルマント周波数の目標値 (T_n) と得られた値 (f_n) の差を f_n で正規化したもの

Table 4　Norm of z_n, the difference between target (T_n) and resultant (f_n) formant frequencies normalized by f_n.

Subject	Vowel				
	/a/	/e/	/i/	/o/	/u/
HT	0.014	0.145	0.013	0.013	0.012
KH	0.014	0.028	0.014	0.013	0.013
SA	0.013	0.019	0.012		
SH	0.013	0.077	0.016	0.013	0.051
TI	0.012	0.014	0.013	0.014	0.013
YT	0.015	0.081	0.014	0.016	0.013

謝辞　本研究は 2009 年 10 月に開催された APSIPA ASC2009 にて発表したものである [14]．本研究で分析した MRI データは，NICT の支援により行われた「人間情報コミュニケーションの研究開発」にて測定されたものである．本研究の一部は，2009 年度総務省 SCOPE (071705001)，2009 年度科研費 (21300071, 21500184, 21330170) の援助を受けた．

文　献

[1] Q. Lin and C. Che, Normalizing the vocal tract length for speaker independent speech recognition, *IEEE Signal Processing Letters*, 2, 201–203 (1995).

[2] H. Wakita, Normalization of vowels by vocal tract length and its application to vowel identification, *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP25-183 (1997).

[3] 江森正, 篠田浩一, 音声認識のための高速最ゆう推定を用いた声道長正規化, 信学論, J83-D-II, 11, 2108–2117 (2000).

[4] 六井淳, 中井満, 下平博, 嵯峨山茂樹, 最尤推定を用いた声道長線形変換による話者正規化, 情報処理学会論文誌, 43, 7, 2030–2037 (2002).

[5] 鏑木時彦, 正木信夫, 元木邦俊, 松崎博季, 北村達也, 音声生成の計算モデルと可視化, コロナ社 (2010).

[6] C.-S. Yang and H. Kasuya, Uniform and non-uniform normalization of vocal tracts measured by MRI across male, female and child subjects, *IEICE Trans. Inf. & Syst.*, E78-D, 732–737 (1995).

[7] 北村達也, 竹本浩典, 本多清志, 母音発声時の声道断面積関数の

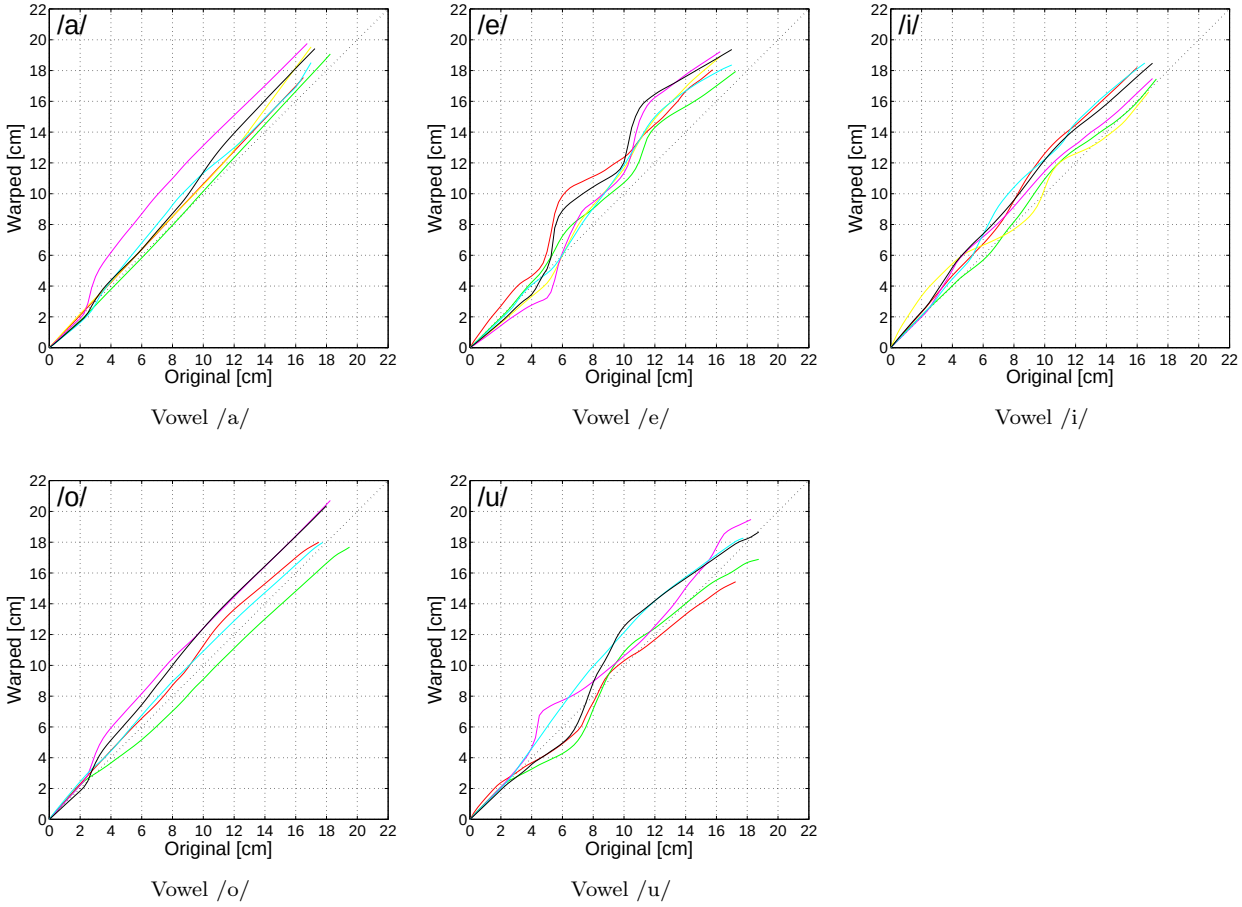

図 5　被験者 6 名の日本語 5 母音の声道ワーピング関数

Fig. 5　Vocal tract warping functions for the five Japanese vowels for six subjects.

個人差について, 音講論 (春), 285–286 (2004).

[8] S. Adachi, H. Takemoto, T. Kitamura, P. Mokhtari and K. Honda, Vocal tract length perturbation and its application to male-female vocal tract shape conversion, *J. Acoust. Soc. Am.*, 121, 3874–3885 (2007).

[9] H. Takemoto, T. Kitamura, H. Nishimoto and K. Honda, A method of tooth superimposition on MRI data for accurate measurement of vocal tract shape and dimensions, *Acoust. Sci. & Tech.*, 25, 468–474 (2004).

[10] H. Takemoto, K. Honda, S. Masaki, Y. Shimada and I. Fujimoto, Measurement of temporal changes in vocal tract area function from 3D cine-MRI data, *J. Acoust. Soc. Am.*, 119, 1037–1049 (2006).

[11] S. Adachi and M. Yamada, An acoustical study of sound production in biphonic singing Xöömij, *J. Acoust. Soc. Am.*, 105, 2920–2932 (1999).

[12] R. Caussé, J. Kergomard and X. Lurton, Input impedance of brass musical instruments – comparison between experiment and numerical models, *J. Acoust. Soc. Am.*, 75, 241–254 (1984).

[13] B.H. Story, Technique for "tuning" vocal tract area functions based on acoustic sensitivity functions, *J. Acoust. Soc. Am.*, 119, 715–718 (2006).

[14] Tatsuya Kitamura, Hironori Takemoto and Seiji Adachi, Vocal tract warping for normalizaing inter-speaker differences in vocal tract transfer functions, *Proc. of APSIPA ASC 2009*, 525–528 (2009).