

音声教育のための声道形の動的3次元表示法

佐々木 優、平野 崇、三輪 譲二

masaru, taka, miwa@cis.iwate-u.ac.jp

岩手大学 工学部 情報工学科

〒020-8551 岩手県盛岡市上田4-3-5
019-621-6974

あらまし 本論文では、音声教育に役立つために、連続音声を発声したときの声道形を動的に3次元表示する方法を検討した。まず、9次元声道形モデルの3次元表示によって、声道形の特徴を視覚的に表現できた。また、調音パラメータの典型値を求めることで、調音—音響領域の対応ができた。さらに、音響次元の特徴に対応する動的声道形を推定した。最後に、推定された動的声道形パラメータによって、声道形の動的3次元表示が可能となった。これにより、連続母音や半母音等の動的調音表現が実現でき、音声発音教育に役立つことができるものと思われる。

キーワード 音声言語教育、声道形、3次元表示、調音モデル、動的表示、調音推定アルゴリズム

Dynamic 3-Dimensional Visualization of Vocal Tract Shape for Speech Learning System

Masaru Sasaki, Takashi Hirano and Jouji Miwa

masaru, taka, miwa@cis.iwate-u.ac.jp

Department of Computer and Information Science,
Faculty of Engineering, Iwate University

4-3-5 Ueda Morioka-shi Iwate-ken, 020-8551 Japan
019-621-6974

Abstract We propose a method of dynamic 3-dimensional visualization of vocal tract shape for computer assisted language learning. In the method, vocal tract shape is modeled 9-dimensional parameters such as $A_g, A_b, A_c, A_f, A_l, X_b, X_c, X_f$ and X_l , the typical values of the articulation parameters are corresponded with "articulatory-acoustic" transformation, dynamic vocal tract parameters corresponding with acoustic features are estimated, and it is possible to visualize dynamically the 3D vocal tract model by estimated articulation. We hope that it is useful for language learning because of the dynamic visualization of connected vowels and semi vowels.

key words Computer assisted language, Vocal tract, 3D visualization, Articulation model, Dynamic visualization, Estimation of articulation

1 序論

音声言語教育では、学習者への効率的な発声援助手段が必要である。従来の音声教育では、学習者の発声した音声を教師が聞き教師の内省などを考慮して、学習者の発声指導を行っているが、CAIシステムでは自動的に学習者の音声を分析し、学習者へフィードバックすることにより発声指導をすることが必要である。

本論文では、音声教育に役立てるため、動的声道形を3次元CGで学習者に表現する方法を提案する。それにより、声道形の特徴を視覚的に理解できるという利点がある。ここで、静的声道形を表示する際に、データ量を圧縮した9次元調音モデルの構築を行う。また、時間的に変化する連続音や半母音等を動的に3次元表示するため、動的な声道形パラメータを推定する方法を提案する。

2 声道形の静的3次元表示

2.1 静的3次元表示

声道の静的3次元表示は、発声中に声道形が時間によって変わらない母音について行なう。なお、声道3次元表示の実現には複数の円柱を連結して近似させる。この円柱の底面積には声道断面積を、高さには声道長をセクション数で割った値を、それぞれ対応させる。ここで、声道断面積はMRIからのデータとして与えられる。そして、底面積がいろいろな大きさの円柱をセクション数分だけ組み合わせて声道を構成する。その静的声道形の3次元表示例を、図1に示す。ここで、図の黒い部分が声道を表している。

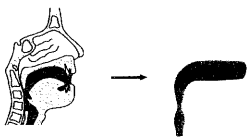


図1: 声道形の3次元表示

2.2 声道断面積関数と伝達スペクトル

音の伝播を平面波と仮定すると、唇での平面出力音圧 P_L と体積速度 U_L は、図2の四端子回路の電圧と電流に対応する [1] [2]。ここで添字の G, L はそれぞれ Glottis, Lip を表している。

声門での入力音圧 P_G と体積速度 U_G から次の関係式で表される。

$$\begin{pmatrix} P_L(\omega) \\ U_L(\omega) \end{pmatrix} = \begin{pmatrix} A(\omega) & B(\omega) \\ C(\omega) & D(\omega) \end{pmatrix} \begin{pmatrix} P_G(\omega) \\ U_G(\omega) \end{pmatrix} \quad (1)$$

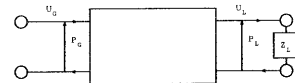


図2: 4端子回路

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} A_1 & B_1 \\ C_1 & D_1 \end{pmatrix} \cdots \begin{pmatrix} A_n & B_n \\ C_n & D_n \end{pmatrix} \quad (2)$$

ここで A, B, C, D はそれぞれ ω の関数である。また、 $B_i(\omega), C_i(\omega)$ は i 番目のセクションの声道断面積 a_i の関数で表わされる。更に、四端子回路の性質から次の関係式が成り立つ。

$$P_L(\omega) = U_L(\omega) \times Z_L(\omega) \quad (3)$$

U_G から P_L までの伝達関数 $H(\omega)$ は、以下で表される。

$$H(\omega) = \frac{P_L(\omega)}{U_G(\omega)} = \frac{Z_L(\omega)}{A(\omega) - C(\omega)Z_L(\omega)} \quad (4)$$

これにより、声道断面積 a_i と伝達関数 $H(\omega)$ の関係が定まり、 $H(\omega)$ からホルマント周波数を求めることができる。

2.3 9次元声道形モデル

ここで、学習者への有効なフィードバック情報として、9次元声道形モデルの構築を行った。MRIから得られた n セクションの声道断面積データを9次元の特徴で簡単化して表現する。この9個のパラメータはそれぞれ意味を持っており、単なる静的3次元表示では表現しづらい中舌化や狭口化を、パラメータの値によって表示することができる。すなわち声道形状を、MRIデータから調音モデルとして、図3に示すように声門面積 A_g 、後室位置 X_b と面積 A_b 、調音位置 X_c と面積 A_c 、前室位置 X_f と面積 A_f 、唇開放面積 A_l 、声道長 X_l の9個のパラメータで簡単化して表現する。声道形は、これらのパラメータをスプライン補間して得る。

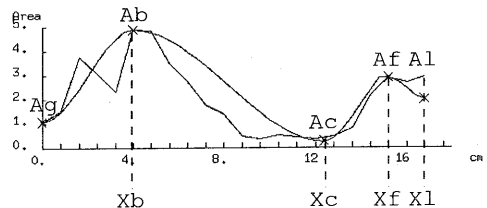


図3: MRIデータと9次元声道形モデル

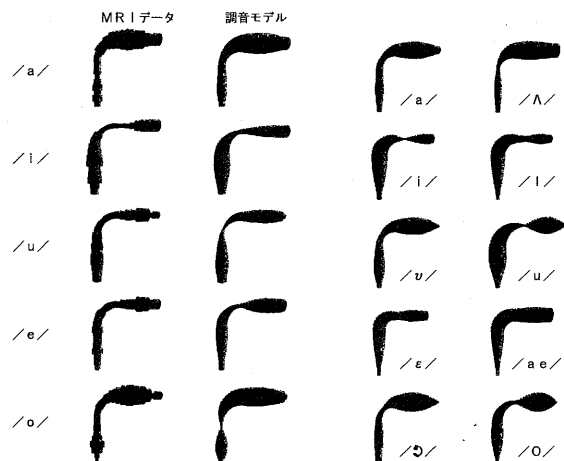


図 5: 日本語母音と英語母音の声道形 3次元表示の例

2.4 日本語母音の声道形モデル

表 1に、成人男性が発声した日本語 5 母音の平均ホルマント周波数 [3] に対応する 9 次元調音モデルによる典型値を示す。

表 1: 日本語母音の調音モデルの典型値

	Area(mm ²)					position(mm)			
	<i>A_g</i>	<i>A_b</i>	<i>A_c</i>	<i>A_f</i>	<i>A_l</i>	<i>X_b</i>	<i>X_c</i>	<i>X_f</i>	<i>X_l</i>
a	67	186	92	840	325	22	68	141	181
i	110	491	37	289	294	44	106	167	181
u	150	308	55	364	98	22	92	155	181
e	48	247	71	422	195	62	97	158	181
o	20	383	66	845	139	22	51	132	181

この表から、/i/ や /e/ は *X_c* が大きいことから前舌母音であり、/a/ は *X_c* が小さいことから後舌母音であることがわかる。調音パラメータ *X_c*、*A_l* を用いた日本語母音の母音五角形を、図 4 に示す。このように

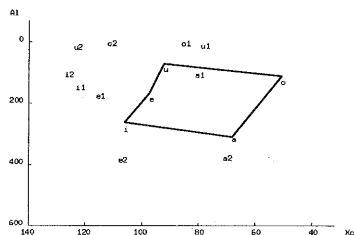


図 4: 日本語と英語母音の *X_c* と *A_l* の調音点

単純化することにより、母音毎の調音位置などの情報を関連付けることができる。

2.5 英語母音の声道形モデル

表 2に、英語 MRI データ [4] から求めた 9 次元声道形モデルによる典型値を示す。また、英語母音の *X_c*、*A_l* 平面図を図 4 に示す。これにより、英語母音は日本語より前舌かつ狭母音傾向があると言える。また表よ

表 2: 英語母音の調音モデルの典型値

	Area(mm ²)					position(mm)			
	<i>A_g</i>	<i>A_b</i>	<i>A_c</i>	<i>A_f</i>	<i>A_l</i>	<i>X_b</i>	<i>X_c</i>	<i>X_f</i>	<i>X_l</i>
Λ	33	122	66	579	117	32	81	139	175
a	45	112	23	655	388	36	71	131	175
i	33	468	10	205	158	67	123	158	167
I	20	396	75	197	116	71	127	151	167
v	33	220	53	569	22	28	79	139	175
u	40	586	15	547	26	48	124	158	183
ε	21	303	136	259	185	71	116	135	159
ae	22	387	319	479	394	87	108	143	167
∅	61	142	32	708	13	36	86	135	175
O	18	272	32	725	14	44	112	147	175

り、/i/ と /I/ では、調音位置はほとんど変わらないが、/I/ の狭めの面積をより小さくすればよい。更に /∅/ と /O/ については逆で、狭めの位置をずらしていることがわかる。

2.6 声道形状の 3次元表示

このようにして求められた 9 次元声道形モデルの典型値を用いて、日本語と英語母音の 3 次元声道形状を構成した [5] [6]。その例を図 5 に示す。図からもわかるように、声道形を静的に 3 次元表示することによって母音を発声したときの声道のイメージがつかめるようになった。更に、各母音の特徴や違いを、視覚的に理解できることがわかる。

表 3: 動的モデル関数の長所、短所

	長所	短所
線形関数	単調関数である	t_1 、 t_2 の前後で不連続である
臨界制動関数	物理モデルに対応する	t_1 の前後で不連続である 適切な r_i を求める必要がある
シグモイド関数	t_1 、 t_2 の前後が比較的連続である 変化量 α_i と時刻 C_i を独立に制御できる	適切な α_i の値を求める必要がある

3 声道形の動的 3 次元表示

3.1 動的モデル

連続母音や半母音等のように時間によって変化する連続音声の動的 3 次元表示を行う。ここで、時間変化に伴う 9 次元声道形パラメータ値 ($A_g, A_b, A_c, A_f, A_l, X_b, X_c, X_f, X_l$) の推移を、関数 $f_i(t)$ を用いて以下のように示す。

$$f_i(t) = \begin{cases} pf_i & (t_0 \leq t \leq t_1) \\ (pb_i - pf_i)g_i(t) + pb_i & (t_1 < t < t_2) \\ pb_i & (t_2 \leq t \leq t_3) \end{cases} \quad (5)$$

$$(i = 1, 2, \dots, 9)$$

ここで、時刻 t_1 は先行音素の終了時刻、 t_2 は後続音素の開始時刻である。また i は、9 次元声道形のパラメータ番号であり、更に pf_i 、 pb_i とは、それぞれ先行、後続音素のパラメータ値を示す。 t_1 から t_2 までの変化度合い関数 $g_i(t)$ は 0 ~ 1 の値を持つ関数で、式 (6) の線形関数モデル、式 (7) の臨界制動関数モデル [7]、式 (9) のシグモイド関数モデルの 3 つの関数を検討する。

$$g_i(t) = \frac{t - t_1}{t_2 - t_1} \quad (6)$$

$$g_i(t) = 1 - (1 - r_i(t - t_1)) \exp(-r_i(t - t_1)) \quad (7)$$

$$g_i(t) = \frac{1}{1 + \exp(-\alpha_i(t - C_i))} \quad (8)$$

$$C_i = \frac{(t_2 - t_1)}{2} \quad (9)$$

この 3 つの関数についての特徴を、表 3 に示す。ここでシグモイド関数に着目すると、このモデルでは、 t_1 、 t_2 両方の前後で比較的連続であり、なおかつパラメータ値の変化が滑らかであるので、現実的にふさわしいモデルである。また、 $t_2 - t_1$ の値により適切な α_i の値を与えてシミュレートすることもできる。そこで、/iia/ と発声したときのシグモイド関数モデルの 9 次元声道形パラメータ変化の例を図 6 に、ホルマント周波

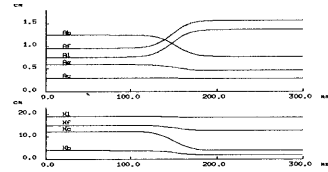


図 6: シグモイド関数モデルのパラメータ変化の例

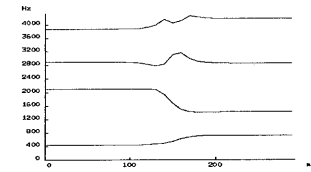


図 7: シグモイド関数モデルの調音音響変換によるホルマント変化の例

数の変化の例を図 7 にそれぞれ示す。図からも、 t_1 、 t_2 両方の前後で比較的連続であるので、現実的にふさわしいモデルであることがわかる。

3.2 動的モデルパラメータ推定アルゴリズム

このシグモイド関数モデルに対し、式 (8) で表わすパラメータ α_i と C_i を推定する。ここで、 α_i と C_i を動的モデルパラメータと呼ぶ。その推定アルゴリズムを、図 8 に示す。この推定アルゴリズムの初段では、 i は一定とし、9 個のパラメータ全てに対して関数 $g_i(t)$ が同一になるようにした。また、分析、調音音響変換による合成ホルマント周波数の軌跡を、それぞれ $F[t]$ 、 $f[t]$ とする。この推定アルゴリズムは、大きく 4 つの過程に分けられる。すなわち C_{fi} 、 α_{fi} 、 C_i 、 α_i の 4 つの変数の推定過程である。

α_i 、 C_i 推定の前にまず、 α_i と $t_2 - t_1$ の関係を調べた。ここで、 $T[\alpha_i] = t_2 - t_1$ と定義する。図 9 のよ

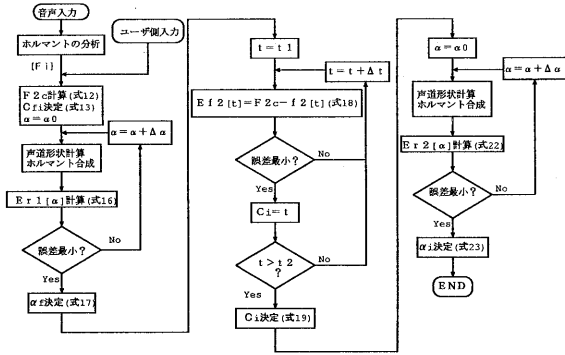


図 8: 動的パラメータ α_i 、 C_i 推定アルゴリズム

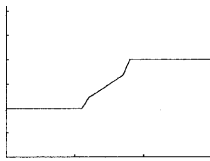


図 9: $T[\alpha_i]$ が小さく不連続となる場合

うに α_i の値に対して $T[\alpha_i]$ が小さ過ぎないように最適な $T[\alpha_i]$ の値を求めた。すなわち、関数 $g_i(t)$ の変化開始から 1 フレーム後の値が 0.02 以下で、なおかつ変化終了から 1 フレーム前の値が 0.98 以上を条件として α_i と $T[\alpha_i]$ の関係を求めた。このようにして得られた α_i と $T[\alpha_i]$ の関係を図 10 に示す。この関係から、 α_i を仮定すると $T[\alpha_i]$ の値を自動的に与えることができる。

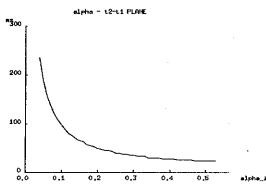


図 10: α_i と $T[\alpha_i]$ の関係

次に、 α_i 、 C_i 推定アルゴリズムを行なう。まず、次式を用いて分析第 2 ホルマント周波数データから $\widehat{F2}_c$ を推定し、これを用いて分析第 2 ホルマントにおける変化の中心時刻 \widehat{C}_{fi} を推定する。

$$\widehat{F2}_f = \frac{\sum_{t=t_{fs}}^{t_{fe}} F_2[t]}{t_{fe} - t_{fs}} \quad (10)$$

$$\overline{F2}_b = \frac{\sum_{t=t_{bs}}^{t_{be}} F_2[t]}{t_{be} - t_{bs}} \quad (11)$$

$$\widehat{F2}_c = \frac{\widehat{F2}_f + \overline{F2}_b}{2} \quad (12)$$

$$\widehat{C}_{fi} = \operatorname{argmin}(\widehat{F2}_c - F_2[C_{fi}]) \quad (13)$$

式 (10)、(11) で先行音素と後続音素の平均ホルマント周波数を求め、式 (12) で求められた $\widehat{F2}_c$ に最も近い分析ホルマント周波数の時刻を \widehat{C}_{fi} と決定する。

このようにして推定された \widehat{C}_{fi} を式 (8) に代入して今度は、 $\widehat{\alpha}_{fi}$ の推定を行なう。図 11 は、様々な値の

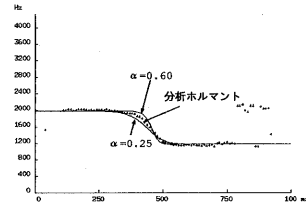


図 11: α の値によるホルマントの変化の例

α に対する合成ホルマントの変化の例である。このようにして、分析ホルマントに対して誤差が最も小さくなるような $\widehat{\alpha}_{fi}$ を推定する。下式は、推定された \widehat{C}_{fi} を用いて $\widehat{\alpha}_{fi}$ を算出する式である。

$$t_1 = \widehat{C}_{fi} - \frac{T[\alpha]}{2} \quad (14)$$

$$t_2 = \widehat{C}_{fi} + \frac{T[\alpha]}{2} \quad (15)$$

$$Er1[\alpha] = \frac{\sum_{t=t_1}^{t_2} (F_2[t] - f_2[t])^2}{t_2 - t_1} \quad (16)$$

$$\widehat{\alpha}_{fi} = \operatorname{argmin}(Er1[\alpha_i]) \quad (17)$$

式 (16) で $t_1 \sim t_2$ の間で 1 フレーム当りの 2 乗誤差を算出し、式 (17) により、その中で最も誤差の小さい α を $\widehat{\alpha}_{fi}$ と決定した。

このようにして推定された \widehat{C}_{fi} 、 $\widehat{\alpha}_{fi}$ に基づいた調音音響変換の結果を図 12 に示す。この図から、調音次元の \widehat{C}_{fi} が必ずしも音響次元においても中心であるとは限らないことがわかった。そこでこの周波数領域でのずれを補正するために、更に式 (12) の $\widehat{F2}_c$ と、 $\widehat{\alpha}_{fi}$ を式 (8) に代入して得られる声道形パラメータによる合成ホルマントの比較から、 \widehat{C}_i を推定した。その式を以下に示す。

$$E_{f_2}[t] = |\widehat{F2}_c - f_2[t]| \quad (18)$$

$$\widehat{C}_i = \operatorname{argmin}(E_{f_2}[C_i]) \quad (19)$$

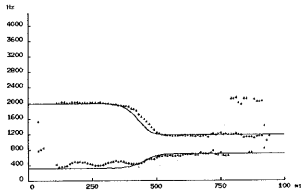


図 12: \widehat{C}_{fi} , $\widehat{\alpha}_{fi}$ に基づいた推定結果 (/iia/)

最後にこの \widehat{C}_i に基づいて $\widehat{\alpha}_i$ の推定を行った。その推定の式を以下に示す。

$$t_1 = \widehat{C}_i - \frac{T[\alpha]}{2} \quad (20)$$

$$t_2 = \widehat{C}_i + \frac{T[\alpha]}{2} \quad (21)$$

$$Er2[\alpha] = \frac{\sum_{t=t_1}^{t_2} (F_2[t] - f_2[t])^2}{t_2 - t_1} \quad (22)$$

$$\widehat{\alpha}_i = \operatorname{argmin}(Er2[\alpha_i]) \quad (23)$$

この式によって推定された $\widehat{\alpha}_i$ に基づいた、調音音響変換の結果を、図 13 に示す。これにより分析データと調音音響変換したデータが非常によく適合していることがわかる。

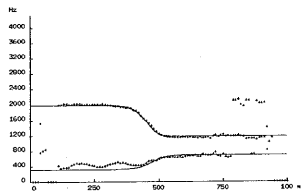


図 13: \widehat{C}_i , $\widehat{\alpha}_i$ に基づいた推定結果 (/iia/)

3.3 動的モデルパラメータの推定結果

前節の推定アルゴリズムを用いて、実音声 /iia/ と /ija/ の推定を行なった。図 14 は、/iia/ と発声したときの音声分析データである。/iia/ と /ija/ を発声した

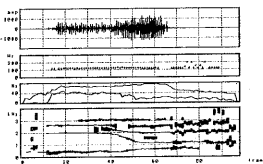


図 14: /iia/ の音声分析結果の例

ときの動的モデルパラメータの推定結果を表 4 に示す。また、その時の 9 次元声道形パラメータの動的モデルパラメータ推定結果を図 15、16 に示す。

表 4: /iia/ と /ija/ の動的モデルパラメータの比較

	$\widehat{\alpha}_i$	$t_2 - t_1$
/iia/	0.38	30 フレーム
/ija/	0.35	28 フレーム

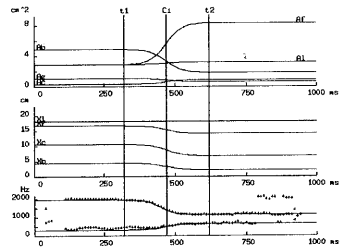


図 15: /iia/ と発声したときの調音音響変換の結果

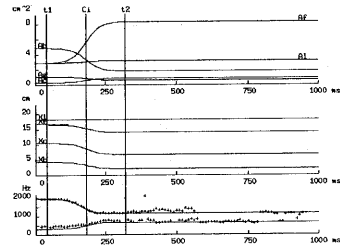


図 16: /ija/ と発声したときの調音音響変換の結果

このように、前節で述べた推定アルゴリズムを用いて、分析ホルマントに対して漸近な合成ホルマントを推定し、その時の 9 次元声道形モデルのパラメータを推定することができた。すなわち、音響次元に対応した動的声道形を推定することができたので、これを用いて 3 次元声道形を表現することができる。なお、今回は、一人の音声データのみを用いたので、推定された $\widehat{\alpha}_i$ の有効性について、今後検討する必要がある。

4 声道形の 3 次元表示

4.1 3 次元 CG 技術の応用

本研究では、3 次元声道形の表示は主に、もっとも一般的であり汎用性の高い OpenGL [8] [9] を用いる。

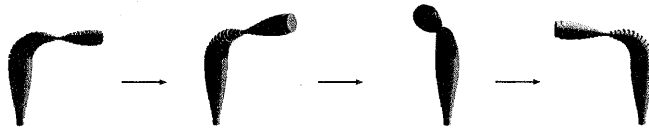


図 17: OpenGL による静的 3 次元表示の回転例

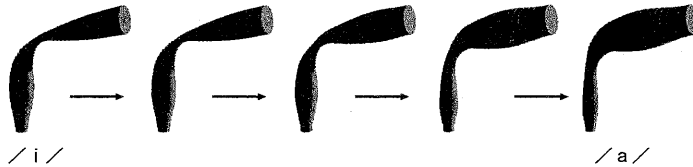


図 18: OpenGL による /ia/ の動的 3 次元表示の例

また、学習者への対話的な表示法として、VRML [10] による声道形状の静的表示も提案する。更に Java を用いて、あらかじめ用意した 3 次元声道形状の画像をアニメーション機能で動的表示するという方法 [11] も提案する。

4.1.1 OpenGL による応用

OpenGL では、表示した声道形を回転させたり拡大したりといった機能や、アニメーション機能が可能なので、声道形の静的表示や動的表示ができる。

OpenGL を使って、声道形を回転させて表示した例を図 17 に示す。回転することにより、調音位置や唇の開き具合などを確認することができる。また、/ia/ と発声したときの声道形を図 18 に示す。これにより、声道形の動的な表示をすることができ、音声教育に利用できることがわかる。

4.1.2 VRML による応用

音声教育への応用としてさらに、VRML による対話型声道形表示を行った。VRML でも声道を、図 19 の

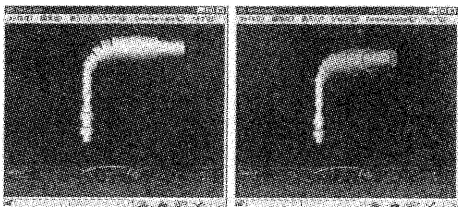


図 19: VRML による声道形の対話型静的表示

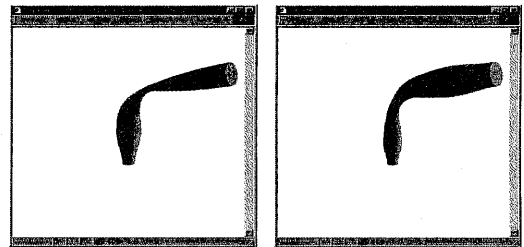


図 20: Java による声道形のアニメーション表示の一部の例

ように回転させたり拡大・縮小ができ、学習者への対話的な表示ができる。しかしながら、声道形の動的 3 次元表示が不可能なため、連続音声の声道形表示はできない。

4.1.3 Java によるアニメーション表示

OpenGL や VRML によってできた 3 次元声道形状を、一定の角度で回転させ、その時の声道形を画像としてファイルに保存する。こうしてできた画像を Java のプログラムに組み込み、アニメーション機能で声道を回転させて表示することができる。その様子を図 20 に示す。このように Java によって、定型の動的 3 次元声道形を表現することができる。しかしこの場合、対話的に声道形を変えることができないという欠点がある。

表 5: 声道形 3 次元表示法の長所・短所

	長所	短所
OpenGL	一般的、対話的で汎用性が高い 動的表現可能	ネットワーク化できない 専用プラットフォーム
VRML	ネットワーク対応 マルチプラットフォーム 設計 (プログラミング) が容易	プラグインが必要 動的表示不可
Java	ネットワーク対応 マルチプラットフォーム 動的表現可能	画像データを用意する必要がある

4.2 3 次元表示の評価検討

表 5 に、これらの技術の長所、短所を示す。今後これらを実際に音声教育に応用し、有効性を検討する必要がある。

5 結論

音声教育に役立てるために、声道形の対話型 3 次元表示法を検討した。

- 声道形の 3 次元表示により、調音位置などの特徴の視覚的な理解が可能となった。
- 音響次元と対応する調音次元の 9 次元声道形モデルの典型値を求めることができた。
- 動的 3 次元表示として、シグモイド関数モデルを用いることにより、時間的に声道形が変化する連母音や半母音等についても 3 次元表示できるようになった。
- 声道形の動的パラメータ α_i 、 C_i の推定アルゴリズムによって、分析データに対して調音パラメータが推定でき、分析データと対応する声道形動的 3 次元表示が可能となった。

今後の課題として、各パラメータに対して独立な α_i 、 C_i の推定を行ない、最適な値を推定する必要がある。また、日本人の発音が苦手な /r/ と /l/ 等の動的 3 次元表示をすることにより、英語教育に役立てることができるものと考えられる。

参考文献

- [1] Man Mouhan Sondhi and Juergen Schroeter : "A Hybrid Time-Frequency Domain Articulatory Speech Synthesizer", IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-35, No. 7, pp. 955-967 (July 1987).
- [2] Jouji Miwa : "Interactive Visualization and Auralization of Speech Production Using Variable Vocal and Nasal Area Function", ASVA97, pp. 271-278 (Apr. 1997).

- [3] 三輪 謙二 : "パソコン音声処理", 昭晃堂, 東京 (July 1991).
- [4] Brad H. Story and Ingo R. Titze : "Vocal tract area functions from magnetic resonance imaging", The Journal of the Acoustical Society of America, Vol. 100, No. 1, pp. 537-554 (July 1996).
- [5] 楊 長盛 : "MR 1 による声道形状の精密計測法の開発と母音の正規化への応用", 宇都宮大学博士論文 (Mar. 1995).
- [6] 佐々木 優, 平野 崇, 三輪 謙二 : "音声教育のための 3 次元声道形状の対話型表現", 日本音響学会春期講演論文集, 2-P-24, pp. 341-342 (Mar. 1998).
- [7] 佐藤 泰雄, 藤崎 博也 : "ホルマント周波数上での調音結合の定式化と音声自動認識への適用", 日本音響学会誌, 34 巻, 3 号, pp. 177-185 (Mar. 1978).
- [8] Jackie Neider, Tom Davis and Mason Woo : "OpenGL Programming Guide", アジソン ウェスレイ (Nov. 1995).
- [9] C. J. Silverio, Beth Fryer and Jed Hartman : "OpenGL Porting Guide", 日本シリコングラフィックス株式会社 (1994).
- [10] 中山 茂 : "Netscape VRML 入門", 日刊工業新聞社, 東京 (July 1996).
- [11] 笹木 望, 太田 晶宏, 藤崎 真美 : "HTML & CGI 入門", エーアイ出版, 東京 (July 1996).