

音声言語教育のための 調音音響変換 A-b-S 法を用いた声道形の推定

平野 崇、三輪 譲二

岩手大学 工学部 情報工学科

あらまし： 本論文では、音声言語教育に役立てるために、調音音響変換による A-b-S(合成による分析)法を用いる声道形の推定法を提案し、その有効性を検討する。声道形モデルとしては、調音位置や唇の開き具合などの声道の特徴を、11 個のパラメータでモデル化した。また、声道形の推定法としては、調音音響変換による A-b-S 法を用いる。ここで、音声言語教育では発音音声既知なので、この情報を声道形を推定するときに利用し、MRI データから求めた典型値を初期値として、声道長を推定した後、A-b-S 法により声道形を推定する。本方法を、日本人が苦手な /l/ を含む実英語音声について適用した結果、動的声道形を推定することが出来、その有効性が確認された。

キーワード： 音声言語教育、声道形推定、調音音響変換、A-b-S、11 次元声道形モデル、典型値

Estimation of Vocal Tract Shape Using A-b-S Method by Articulatory-Acoustic Transformation for Spoken Language Learning.

Takashi Hirano and Jouji Miwa

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

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

Abstract: This paper describes an estimation system of the vocal tract shape using A-b-S (Analysis-by-Synthesis) method for computer assisted spoken language learning (CALL). In the analysis stage of the method, learner's formant frequencies are extracted. In the synthesis stage, 11 parameters such as places and areas of articulation are used as a model of vocal tract shape, initial value of known phoneme is set typical value required from the MRI data, and synthesized parameters are calculated with articulatory-acoustic transformation. In the last stage, the vocal tract shape is estimated with minimum error function. An estimation experiment is carried out for real English speech uttered by a native speaker such as /la/ and /ala/ which are difficult to pronounce for Japanese. From the experiment, the dynamic vocal tract shape is correctly estimated and the effectiveness was confirmed.

Key words : Spoken language learning, Estimation of vocal tract shape, Articulatory-acoustic transformation, Analysis by synthesis, 11 dimensional vocal tract model, Typical value

1 はじめに

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。それに対し、教師のいない学習環境においては、CAIシステムを用いることで、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を行い、発声指導をすることが必要である。

そこで、本論文では、音声言語教育に役立てるため、調音音響変換 A-b-S (Analysis-by-Synthesis: 合成による分析) 法を用いた声道形の推定法を提案し、実音声によりその有効性を検討する。声道形モデルとしては、我々が今まで用いていた 9 次元声道形モデル [1][2][3] を拡張した 11 次元声道形モデルを構築する。この声道形モデルは、調音位置や唇の開き具合などの声道の特徴を 11 個のパラメータでモデル化したものである。また、声道形の推定法としては、調音音響変換 A-b-S 法を用いる。A-b-S 法では、解の収束性や初期値といった問題がある。これらの問題を、我々は、音声言語教育では発音音声は既知であるという情報を利用し、MRI データから求めた典型値を用いることにより、解決した。

図 1 に、音声言語教育システムの構成図を示す。すなわち、音声を学習者からマイクを通して受け取り、その音声に対して音響特徴の抽出を行い、声道形を推定して学習者へフィードバックすることにより音声言語教育に役立てる。

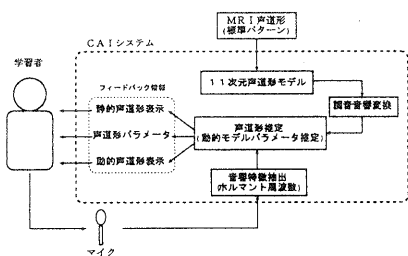


図 1: 音声言語教育システムの構成図

2 11次元声道形モデル

2.1 11次元声道形モデル

我々は、今まで 9 次元声道形モデルを用いて日本語 5 母音の声道形の推定 [3] などを行ってきた。しかしながら、本研究で声道形を推定するとき必要となる /r/ や /l/ の MRI データ [4] を 9 次元声道形モデルに適応させると不都合が生じた。そこで、9 次元声道形モデルを拡張した 11 次元声道形モデルを構築した。ここで、11 次元声道形モデルとは、図 2 に示すように MRI データから、声門面積 A_g 、後室位置 X_b とその断面積 A_b 、後ろ側の狭めの位置 X_{cb} とその断面積 A_{cb} 、前室位置 X_f とその断面積 A_f 、前側の狭めの位置 X_{cf} とその断面積 A_{cf} 、声道長 X_l と唇開放面積 A_l の 11 個の特徴を、それぞれの点が頂点となるような制限付きスプライン補間をし、表現したものである。

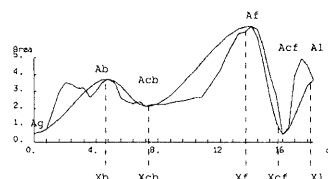


図 2: /l/ の MRI データと 11 次元声道形モデル

2.2 調音音響変換

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、四端子回路の電圧と電流に対応 [5][6] する。ここで添字の G, L はそれぞれ Glottal, 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。

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

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

$$H(\omega) = \frac{P_L(\omega)}{U_G(\omega)} = \frac{Z_L(\omega)}{A(\omega) - C(\omega)Z_L(\omega)} \quad (2)$$

以上のことから、声道断面積 a_i と伝達関数 $H(\omega)$ の関係が定まり、ホルマント周波数を求めることができる。図 3 に、この対応関係の例を示す。

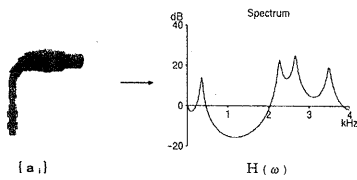


図 3: 声道形と伝達関数の対応関係の例

2.3 ホルマント周波数算出の精度

表 1 に、MRI データ [4] と 11 次元声道形モデルから調音音響変換により算出したホルマント周波数の相対誤差を示す。この結果より、/r/ の F2 で 16.2% と誤差が大きいところもあるが、それ以外では 10% 以内に収まっているので、約 45 セクションのデータを 4 分の 1 の 11 個のパラメータによる声道形モデルで良く近似出来ていることがわかる。

表 1: MRI と声道形モデルによる F1-F3 の相対誤差

	セクション数	F1	F2	F3
/l/	46	6.1%	4.1%	9.8%
/r/	44	6.3%	16.2%	4.0%
/a/	44	9.1%	7.5%	0.9%

2.4 英語子音 /r/、/l/ 及び /a/ の典型値

表 2 に、MRI データ [4] から求めた英語子音 /r/、/l/ 及び /a/ の 11 次元声道形モデルによる声道形パラメータの典型値を示す。ここで、断面積の単位は mm²、調音位置の単位は mm である。

この表の値は、声道形を推定する時の初期値として用いることにより計算量を軽減させている。

図 4 に、英語子音 /r/ と /l/ の声道形 3 次元表示を示す。

2.5 調音次元と音響次元の対応

子音 /l/ の声道形パラメータ X_{cb} 、 A_{cb} 、 X_{cf} 、 A_{cf} 及び Al をそれぞれ独立に変化させたときの

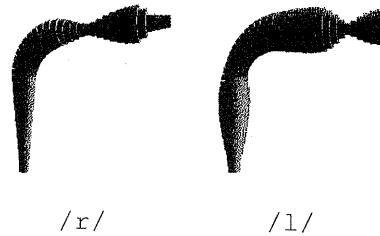


図 4: 英語子音 /r/ と /l/ の声道形 3 次元表示

声道形状を図 5 に示す。また、そのとき音響次元でどのような変化が起こるか、すなわちホルマント周波数がどのように変化するかを調べた。そのときの F1-F2 平面、F1-F3 平面での様子をそれぞれ図 6、図 7 に示す。

この結果、音響次元と調音次元の対応関係が明らかになり、声道形推定のための戦略情報となる。

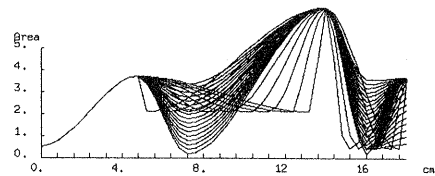


図 5: /l/ の声道形モデル X_{cb} 、 A_{cb} 、 X_{cf} 、 A_{cf} 、 Al の変化

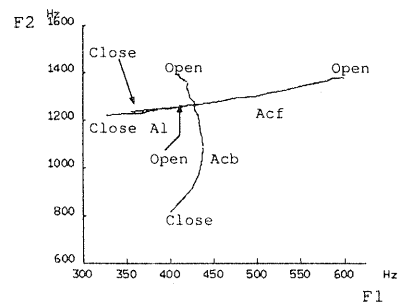


図 6: /l/ の声道形パラメータと F1-F2 平面の関係

表 2: 英語子音 /r/、/l/ 及び /a/ の声道形パラメータの典型値

		A_g	A_b	A_{cb}	A_f	A_{cf}	A_l	X_b	X_{cb}	X_f	X_{cf}	X_l
典型値	/a/	45	112	23	655	387	503	32	56	128	152	172
	/r/	40	310	44	360	67	40	100	128	156	164	172
	/l/	55	376	213	691	45	370	48	72	140	160	180
推定値	/l/	55	376	260	691	100	400	53	70	154	176	198

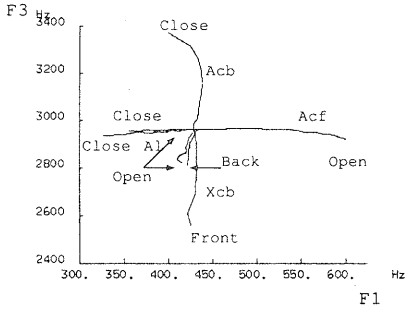


図 7: /l/ の声道形パラメータと F1-F3 平面の関係

3 声道形の推定法

3.1 音声データ

本研究では、日本人が英語を学習するときに特に苦手とされる /r/ や /l/ の子音を含む連続音を、英語を母国語とする成人男性 1 名の発音した音声を使用した。

その音声分析条件は次の通りである。

- (1) サンプリング周波数: 12kHz
- (2) 分析窓: 30ms ハミング窓
- (3) 分析方法: 14 次 LPC
- (4) ホルマント抽出法: ピークピッキング法

ただし、実音声を分析した結果、ホルマント周波数に連続性が得られなかった部分は、連続性を得るために手動で分析結果を修正した。

3.2 声道形動的モデル

連続音や半母音など時間的に連続な音素は、時間の推移によって声道形も変化する。これらの時間によって変化する連続音声を各フレーム毎に任

意に声道形を推定すると声道形の連続性が保たれなくなる可能性があるので、声道形動的モデル [2] を導入し声道形の連続性を保つようにした。

そして、その動的モデルの構成には、関数 $f_i(t)$ を用いて以下のように示す。

$$f_i(t) = \begin{cases} p_{f_i} & (t_0 \leq t \leq t_1) \\ (p_{b_i} - p_{f_i})g_i(t) + p_{b_i} & (t_1 < t < t_2) \\ p_{b_i} & (t_2 \leq t \leq t_3) \end{cases} \quad (3)$$

$(i = 1, 2, \dots, 11)$

ここで、時間 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、11 次元声道形のパラメータ番号であり、さらに p_{f_i} 、 p_{b_i} 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 の値を持つ関数であり、本研究ではシグモイド関数モデルを使用するので、関数 $g_i(t)$ は次のように表せる。

$$g_i(t) = \frac{1}{1 + \exp(-\alpha_i(t - c_i))} \quad (4)$$

$$c_i = \frac{t_1 + t_2}{2} \quad (5)$$

ここで、 α_i と c_i を声道形動的モデルパラメータと呼ぶ。

3.3 推定アルゴリズム

図 8 に、調音音響変換 A-b-S 法を用いた声道形を推定する流れ図を示す。すなわち、音声を分析してホルマント周波数を求め、また、本研究では、音声言語教育ということであらかじめ発音音声が既知であるという情報を利用し、MRI データから求めた典型値を初期値とする。そして、それを調音音響変換により合成してホルマント周波数を求め、音声分析して求めたホルマント周波数との誤差を計算する。さらに、その誤差が最小になるように、声道形を修正していく。

声道形の推定方法は、次の3段階で行う。

1. 声道長推定 (γ)(母音部)
2. 声道形パラメータ推定 (子音部)
3. 動的モデルパラメータ推定 (α_i, c_i)(母音、子音部)

また、動的モデルパラメータ α_i, c_i の推定には、佐々木らの推定アルゴリズム [2] を用いた。

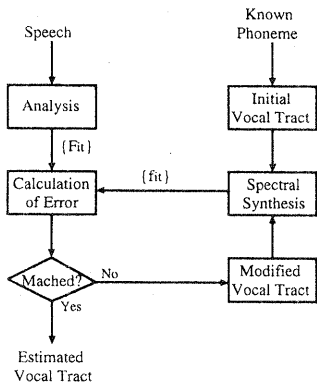


図 8: 言語学習のための A-b-S 型声道形推定

3.3.1 声道長推定方法

身長に個人差があるように声道長にも個人差がある。この個人差を考慮するために、声道長の倍率 γ [3][7] を導入した。

この声道長の推定方法は、 γ の値に応じて Xb, Xcb, Xf, Xcf 及び Xl を平行移動するようにする。本研究では、 γ の範囲を 0.90 から 1.10 の 0.02 刻みとした。また、声道長の推定は、ホルマント周波数が安定している母音の部分で行なった。

3.3.2 声道形パラメータ推定方法

図 6、図 7 から表 3 に示すように、声道形パラメータは、 Acf, Acb, Xcb, Al, Acb の順で推定する。すなわち、音声の子音部分のホルマント周波数の平均を求め、その平均値に近づくように表 3 の条件を利用する。ここで、表 3 の断面積の単位は mm^2 、調音位置の単位は mm である。この

修正条件を、順番ごとに総当たりで行って誤差 E が最小のときの声道形パラメータを見つける。

表 3: 声道形パラメータ修正条件

推定順	声道形パラメータ	刻み幅	最小値	最大値	高感度 F_i
1	Acf	10	10	100	$F1$
2	Acb	10	160	260	$F2$
3	Xcb	4	80	120	$F3$
4	Al	10	320	420	$F1$
5	Acb	10	160	260	$F3$

$$E = \sum_{i=1}^3 (f_i - \bar{F}_i)^2 \quad (6)$$

ここで、 f_i は、声道形パラメータを修正して得られる合成第 i ホルマント周波数であり、 \bar{F}_i は、母音又は子音部分の平均の分析第 i ホルマント周波数である。

3.4 推定結果

前節の推定アルゴリズムを用いて、本研究では、米国人成人男性 1 名による実音声 /la/ と /ala/ の推定を行った。声道長の正規化を行った結果、/la/、/ala/ それぞれの γ の値は、1.10、1.08 であった。図 9 と図 10 に、/la/ と発音したときの音響次元と調音次元での推定結果を示し、図 11 と図 12 に、/ala/ と発音したときの音響次元と調音次元での推定結果を示す。また、/la/ の /l/ の推定値を表 2 の下に示す。

それぞれの音響次元での推定結果を見てみると、/ala/ では $F1$ から $F3$ のすべてのホルマント周波数である程度精度良く推定されたが、/la/ では $F1, F3$ はある程度精度良く推定されたが、 $F2$ では推定誤差が大きく生じた。これは、典型値のズレが影響したものと思われる。また、/ala/ の推定結果の $F2$ では、/a/ から /l/ または /l/ から /a/ に変化するとき単一シグモイド関数では起こらない波のような変化が起きた。これは、本研究では、声道形モデルの 11 個のパラメータすべての動的モデルパラメータ α_i, c_i の値を同じにしているためだと考えられ、今後の検討を必要とする。

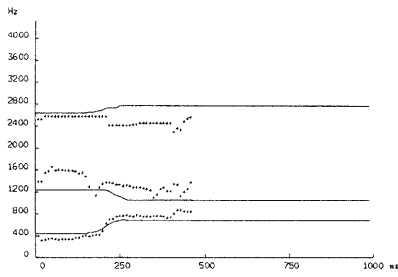


図 9: /la/の音響次元の推定結果

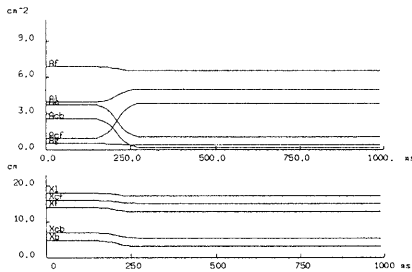


図 10: /la/の調音次元の推定結果

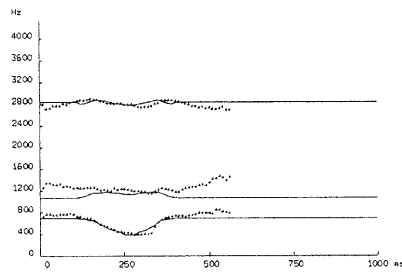


図 11: /ala/の音響次元の推定結果

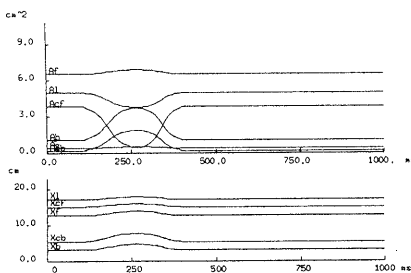


図 12: /ala/の調音次元の推定結果

4 まとめ

本論文では、音声言語教育に役立てるために、調音音響変換 A-b-S 法を用いて声道形の推定を行い、その有効性を検討した。

本研究により得られた結果の要約を以下に示す。

- (1) 声道形を推定するために、我々が今まで用いていた 9 次元声道形モデル [1][2][3] を拡張した 11 次元声道形モデルを構築した。
- (2) 調音次元と音響次元との対応関係が明らかになった。
- (3) 声道長の推定を行うことにより、声道長の個人差を考慮することが可能となった。
- (4) 日本人が英語を学習するときに特に困難とされる /r/ や /l/ の入った子音について、動的声道形状を推定できた。

今後の課題としては、声道形パラメータ推定のアルゴリズムの改良、調音結合の考慮などを検討していく必要がある。

参考文献

- [1] 佐々木 優, 平野 崇, 三輪 譲二: "音声教育のための 3 次元声道形状の対話型表現", 日本音響学会春季講演論文集, 2-P-24, pp.341-342 (Mar. 1998).
- [2] 佐々木 優, 平野 崇, 三輪 譲二: "音声教育のための声道形の動的 3 次元表示法", 電子情報通信学会技術研究報告, SP98-140, pp.31-38 (Feb.1999).
- [3] 平野 崇, 三輪 譲二: "調音-音響変換を用いた 9 次元声道形状の推定", 日本音響学会春季講演論文集, 2-P-16, pp.299-300 (Mar. 1999).
- [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] 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).
- [6] Jouji Miwa: "Interactive Visualization and Auralization of Speech Production Using Variable Vocal and Nasal Area Function", ASVA97, pp.271-278 (Apr. 1997).
- [7] 大村 浩, 田中和世: "ホルマント解析における声道パラメータモデルの話者適応化", 日本音響学会秋季講演論文集, 1-2-13, pp.221-222 (Sep. 1997).