音声合成の手法

kino (http://kinon.sakura.ne.jp/)

2014 年 1 月 31 日

1　時間-周波数領域での音声合成

Sondhi and Schroeter [1999] に基づいて声帯の二質量モデルと声道の音響フィルタによる音声合成システムを記述する。

x_1, x_2 及び u_g、V はそれぞれ声帯質量の変位、声門を通る体積流と声道入口における圧力を表す。

1.1　声道

声道を直径の異なる円筒の接続によりモデル化する。円筒の伝達特性は、入力される圧力 P_{in} 及び体積流 U_{in} と出力する圧力 P_{out} 及び体積流 U_{out}(大文字により周波数領域の函数である事を示している) の関係式

$$\begin{pmatrix} P_{\mathrm{out}} \\ U_{\mathrm{out}} \end{pmatrix} = \begin{pmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{pmatrix} \begin{pmatrix} P_{\mathrm{in}} \\ U_{\mathrm{in}} \end{pmatrix} = K \begin{pmatrix} P_{\mathrm{in}} \\ U_{\mathrm{in}} \end{pmatrix}$$

で表す事が出来、長さ Δl の円筒について

$$K = \begin{pmatrix} \cosh(\sigma \Delta l) & -\frac{1}{\beta}\sinh(\sigma \Delta l) \\ -\beta \sinh(\sigma \Delta l) & \cosh(\sigma \Delta l) \end{pmatrix} \tag{1}$$

となっている。ここで

$$\sigma^2 = j\omega \left(\frac{j\omega}{c^2} + \frac{\rho}{A(x)} Y(s) \right)$$
$$\beta = \sqrt{\frac{A(x)}{\rho} \left(\frac{Y(\omega)}{j\omega} + \frac{A(x)}{\rho c^2} \right)}$$

であり、Sondhi [1974] によると

$$Y(\omega) = \frac{\omega_0^2 j\omega}{(j\omega + a)j\omega + b} + \sqrt{c_1 j\omega}$$

である[*1]。

以下 Sondhi and Schroeter [1987] による具体的な伝達特性の計算について説明する。軟口蓋は声門の膨張の下流 8cm にあるとし、狭めは軟口蓋から口唇の間でのみ起こるとする。

[*1] Sondhi and Schroeter [1987] の R にあたる効果は $\sqrt{c_1 s}$ に含まれるという。

表 1 声道モデルのパラメータ

パラメータ	内容	値	単位
c	音速	3.5×10^4	cm/s
ρ	空気密度	1.14×10^3	g/cm^3
	(口腔)		
Δl	声道を近似する円筒要素の長さ	0.85	cm
a	壁面のレジスタンスの質量比	130π	rad/s
b	機械的共振の二乗角振動数	$(30\pi)^2$	(rad/s)^2
c_1	熱伝導率及び粘性係数の補正	4	rad/s
ω_0^2	音響的共振の最低二乗角振動数	$(406\pi)^2$	(rad/s)^2
	(鼻腔。その他の変数は全て口腔と同じ)		
c_1	熱伝導率及び粘性係数の補正	72	rad/s
	(副鼻腔)		
$R_{\sin}$	結合部位の音響抵抗	1	dyn $\cdot$ s/cm^5
$L_{\sin}$	結合部位の音響リアクタンス	5.94×10^{-3}	g/cm^4
$C_{\sin}$	20.8 cm^3 の音響コンプライアンス	15.8×10^{-6}	cm$^4 \cdot$ s^2/g

1.1.1 口腔

声道の異なる部分は、K_G が声門から軟口蓋、K_N が軟口蓋から鼻孔、K_C が軟口蓋から狭め、K_L が狭めから口唇の間という 4 つの連鎖行列により表現される。K_G, K_N, K_C, K_L の各連鎖行列は、(1) により得られる区間毎の行列を合成して得られるが、軟口蓋のところで少し工夫が要る。声門から口唇に至る連鎖行列を求めるには、鼻腔の方の枝を

$$K_{cN} = \begin{pmatrix} 1 & 0 \\ -1/Z_{VN} & 1 \end{pmatrix} \tag{2}$$

と表現する。ここで Z_{VN} は軟口蓋に於ける鼻腔の入力インピーダンスである。声門から鼻孔に至る連鎖行列を求めるには、口腔側の入力インピーダンス Z_{VT} により同様な行列 K_{cT} を求める。

1.1.2 鼻腔

鼻腔は 11cm の縦続音響管でモデル化されるが、この場合伝達函数の最初の極零対が高過ぎる周波数を持つ。そこで副鼻腔を、軟口蓋から 7cm のところで鼻腔に接続した

$$Z_{\sin} = R_{\sin} + j\omega L_{\sin} + \frac{1}{j\omega C_{\sin}}$$

なるインピーダンスの Helmholtz 共鳴器によりモデル化すると、より実際に近い値が得られる。このインピーダンスは鼻腔の連鎖行列へ (2) と同様にして組み込まれる。

鼻腔の形状は Maeda [1982] による Table 2 on page 3 のデータを用いる。

表 2 鼻腔の断面積函数

軟口蓋からの距離 (cm)	断面積 (cm^2)
0	1
1	2
2	3
3	4
4	6
5	8
6	8
7	7
8	4
9	2
10	2

1.1.3 時間領域での音声合成

以上に定義した連鎖行列 $K_G, K_N, K_{C,}K_L, K_{cN}, K_{cT}$ が得られたとして、モデル全体を表現する行列を導く。声門から狭めまでの連鎖行列は

$$K_{\text{fric}} = K_C K_{cN} K_G$$

で、声門から口唇までの連鎖行列は

$$K_{\text{tract}} = K_L K_{\text{fric}}$$

で与えられる。また声門から鼻孔までは

$$K_{\text{nasal}} = K_N K_{cT} K_G$$

となる。声道全体の入力インピーダンスは

$$Z_{in} = \frac{k_{22}^{\text{tract}} Z_L - k_{12}^{\text{tract}}}{k_{11}^{\text{tract}} - k_{21}^{\text{tract}} Z_L}$$

であり、Z_L は口唇での放射インピーダンスを表す。Z_{VN} や Z_{VT} も同様な式である。放射インピーダンスは口唇の開きと同じ半径 r_L の呼吸球と等しく

$$Z_L = \frac{4\pi r_L^2 \rho c}{c^2 + r_L^2 \omega^2}(r_L^2 \omega^2 + c r_L j\omega)$$

で求められる[*2]。

[*2] Z_{VN} で用いられるであろう Z_N は？

U_g 及び P_L で声門体積流 u_g 及び口唇から放射される音圧 p_L の Fourier 変換を表す。U_g から P_L への伝達函数は

$$H_L = \frac{P_L}{U_g} = \frac{Z_L}{k_{11}^{\mathrm{tract}} - k_{21}^{\mathrm{tract}} Z_L}$$

であり、閉じた声道に於いては 0 となる。また U_g から P_L への伝達函数は

$$H_N = \frac{P_N}{U_g} = \frac{Z_N}{k_{11}^{\mathrm{nasal}} - k_{21}^{\mathrm{nasal}} Z_N}$$

で、鼻腔が声道に接続されていない場合 (接続面積 $A_{\mathrm{coupl}} = 0$) は 0 となる。有声音の出力は、伝達函数

$$H_{\mathrm{out,voiced}} = \frac{P_{\mathrm{speech}}}{U_g} = H_L + H_N + H_{\mathrm{vib}}$$

の逆 Fourier 変換により得られるインパルス応答 h_{out} と u_g の畳み込みで得られる。ここで伝達函数

$$H_{\mathrm{vib}} = \frac{A_1}{c} \cdot \frac{j\omega r_{\mathrm{vib}}}{c + j\omega r_{\mathrm{vib}}} Z_{\mathrm{in}} \beta$$

は、声門に於ける声道壁の粒子速度で振動する半径 r_{vib} の球から放射される音圧を表す。A_1 は声門側の声道断面積の最初の値で、Z_{in} は同じ平面での声道の入力インピーダンスである。

声道の伝達特性は 20ms 毎に計算され、再計算までの間はそれを線形補完した値が用いられる。

$$\begin{aligned} V(t) &= z_{in}(t) * u_g(t) \\ &\approx \int_0^T z_{in}(\tau) u_g(t-\tau) d\tau \end{aligned}$$

が現在の V を与える。$Z_{in}(\omega)$ のインパルス応答 $z_{in}(t)$ は

$$z_{in}(n) \approx \sum_{m=0}^{N} c_f(m) c_t(n) |Z_{in}(2\pi m f_0)| \cos(2\pi m f_0 n \Delta t)$$

と近似される。ここで c_f, c_t は高周波・長時間の成分を減衰させる係数であり

$$\begin{aligned} c_f(m) &= \left[1 + \exp\left\{ 4 \left(\frac{m}{N} - \frac{5}{8} \right) \Big/ \frac{1}{8} \right\} \right]^{-1} \\ c_t(n) &= 0.54 - 0.46 \cos\left(\pi \frac{N+m}{2N} \right) \end{aligned}$$

の様に取れる[*3]。また周波数分解能は $N\Delta t = T = 1/f_0$ の関係によって決定される。

1.2 声帯

カオス性は取り入れていないが、Koga and Nakagawa [1998] の記述を参考に説明する。

[*3] c_t は元論文の通り hamming 窓の右半分である。c_f については「1/2 から 3/4 にかけて減衰する」という記述を満たす様にデザインした。

表 3 声帯モデルのパラメータ

パラメータ	内容	値	単位
A_{g0}	静止状態での声門の隙間の断面積	0.05	cm^2
l_g	声帯の有効長 (声門の隙間)	1.4	cm
x_c	声門の衝突時の変位	$-A_{g0}/2l_g$	cm
P_s	声門下圧	7850	$\mathrm{dyn/cm}^2$
m_1	下側の声帯質量	$0.125/q$	g
m_2	上側の声帯質量	$0.025/q$	g
d_1	m_1 の厚さ	$0.25/q$	cm
d_2	m_2 の厚さ	$0.05/q$	cm
η_{k1}	非線形発条定数	100	-
η_{k2}	非線形発条定数	100	-
η_{h1}	非線形発条定数	500	-
η_{h2}	非線形発条定数	500	-
k_{s1}	線形発条定数	$80000q$	dyn/cm
k_{s2}	線形発条定数	$8000q$	dyn/cm
k_{h1}	非線形発条定数	$3k_1$	dyn/cm
k_{h2}	非線形発条定数	$3k_2$	dyn/cm
k_c	結合発条係数	$25000q^2$	dyn/cm
μ	空気の粘性係数	1.86×10^{-4}	$\mathrm{dyn \cdot s/cm}^2$
ρ	空気の密度	1.14×10^{-3}	$\mathrm{g/cm}^3$
$r_{1_{\mathrm{open}}}$	ダンピング抵抗[*1]	$2 \times 0.2\sqrt{k_1 m_1}$	g/s
$r_{1_{\mathrm{closed}}}$	ダンピング抵抗[*1]	$2 \times 1.1\sqrt{k_2 m_2}$	g/s
$r_{2_{\mathrm{open}}}$	ダンピング抵抗[*1]	$2 \times 0.6\sqrt{k_1 m_1}$	g/s
$r_{2_{\mathrm{closed}}}$	ダンピング抵抗[*1]	$2 \times 1.9\sqrt{k_2 m_2}$	g/s

1.2.1 流体系

声門における圧力減衰は

$$V - P_s = -(R_{v1} + R_c + R_{12} + R_e + R_{v2})u_g - (L_{g1} + L_{g2})\dot{u}_g$$

であり、Koga and Nakagawa [1998] によれば

$$R_{vi} = \frac{3\mu}{2l_g} \cdot \frac{d_i}{h_i^3},\ L_{gi} = \frac{\rho}{2l_g} \cdot \frac{d_i}{h_i}$$

$$R_c = \frac{\rho}{8l_g^2} \cdot \frac{1.37}{h_1^2}|u_g|,\ R_{12} = \frac{\rho}{8l_g^2}\left(\frac{1}{h_2^2} - \frac{1}{h_1^2}\right)|u_g|,\ R_e = -\frac{\rho}{8l_g^2} \cdot \frac{0.5}{h_2^2}|u_g|$$

となっている。ここで $h_i = x_i - x_c$ である。前節の様に V を求めれば、これは u_g の一階微分方程式となる。

1.2.2 機械系

声帯質量の運動方程式は

$$\begin{cases} m_1\ddot{x}_1 + r_1\dot{x}_1 + s_1(x_1) = P_1 l_g d_1 - k_c(x_1 - x_2) \\ m_2\ddot{x}_2 + r_2\dot{x}_2 + s_2(x_2) = P_2 l_g d_2 - k_c(x_2 - x_1) \end{cases}$$

であり、ここで

$$\begin{aligned} P_{11} - P_s &= -R_c u_g \\ P_{12} - P_{11} &= -R_{v1} u_g - L_{g1}\dot{u_g} \\ P_{21} - P_{12} &= -R_{12} u_g \\ P_{22} - P_{21} &= -R_{v2} u_g - L_{g2}\dot{u_g} \\ V - P_{22} &= -R_e u_g \end{aligned}$$

より

$$\begin{aligned} P_1 &= \frac{P_{11} + P_{12}}{2} = P_s - \left(R_c + \frac{R_{v1}}{2}\right) u_g - \frac{L_{g1}}{2}\dot{u_g} \\ P_2 &= \frac{P_{21} + P_{22}}{2} = P_1 - \left(R_c + \frac{R_{v1} + R_{v2}}{2}\right) u_g - \frac{L_{g1} + L_{g2}}{2}\dot{u_g} \\ &\stackrel{?}{=} V + \left(R_e + \frac{R_{v2}}{2}\right) u_g + \frac{L_{g2}}{2}\dot{u_g} \end{aligned}$$

また Ishizaka and Flanagan [1972] によれば

$$\begin{aligned} s_i(x_i) &= f_{si}(x_i) + f_{hi}(x_i) \\ &= k_{si} x_i (1 + \eta_{si} x_i^2) + k_{hi} h_i (1 + \eta_{hi} h_i^2) \cdot \theta(-h_i) \end{aligned}$$

である。

1.2.3 声門閉鎖時

$h_1 \leq 0$ または $h_2 \leq 0$ の場合は声門が閉じているから

$$u_g = 0$$

が成り立たなくてはならない。 また声道内の圧力は

$$\begin{aligned} P_1 &= P_s \\ P_2 &= \begin{cases} P_s & (h_1 > 0) \\ 0 & (h_1 \leq 0) \end{cases} \end{aligned}$$

とする。

1.3 全体系

以上より u_g, x_1, x_2 に関する連立微分方程式

$$\begin{cases} (L_{g1}+L_{g2})\dot{u_g} + (R_{v1}+R_c+R_{12}+R_e+R_{v2})u_g + z_{in} * u_g - P_s = 0 \\ m_1\ddot{x_1} + r_1\dot{x_1} + s_1(x_1) + k_c(x_1 - x_2) - P_1 l_g d_1 = 0 \\ m_2\ddot{x_2} + r_2\dot{x_2} + s_2(x_2) + k_c(x_2 - x_1) - P_2 l_g d_2 = 0 \end{cases}$$

が得られる。

参考文献

K. Ishizaka and J. L. Flanagan. Synthesis of voiced sounds from a two-mass model of the vocal cords. *Bell System Tech. J.*, 51(6):1233–1268, 1972.

Hiroyuki Koga and Masahiro Nakagawa. A synthesis model of chaotic vocal sounds. *IEICE Technical Report*, 10:25–32, 1998.

Shinji Maeda. The role of the sinus cavities in the production of nasal vowels. *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '82*, 7:911–914, 1982.

M. M. Sondhi. Model for wave production in a lossy vocal tract. *J. Acoust. Soc. Am.*, 55(5):1070–1075, 1974.

M. M. Sondhi and J. Schroeter. Speech production models and their digital implementations. In Vijay K. Madisetti and Douglas B. Williams, editors, *Digital Signal Processing Handbook*, chapter 44. CRC Press LLC, 1999.

Man Mohan Sondhi and Juergen Schroeter. A hybrid time-frequency domain articulatory speech synthesizer. *IEEE Trans. Acoust., Speech, Signal Process.*, ASSP-35(7):955–967, July 1987.