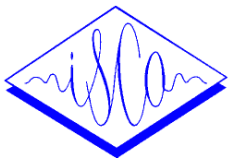

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

IMPROVING THE KELLY–LOCHBAUM VOCAL TRACT MODEL USING CONICAL TUBE SECTIONS AND FRACTIONAL DELAY FILTERING TECHNIQUES

Vesa Välimäki and *Matti Karjalainen*

Helsinki University of Technology, Acoustics Laboratory
Otakaari 5A, FIN-02150 Espoo, Finland

ABSTRACT

An articulatory model of speech production is usually constructed by approximating the profile of the vocal tract using cylindrical tube sections. This is implemented by a digital ladder filter that is called the Kelly–Lochbaum model. In this paper we propose an extended approach, where the tube sections approximating the profile of the tract are conical instead of cylindrical. Furthermore, the length of each tube section in our model can be accurately controlled using a novel fractional delay filtering scheme. These refinements result in an accurate and intuitively controllable vocal tract model that is well suited for articulatory speech synthesis.

I. INTRODUCTION

An articulatory model for the human vocal tract is traditionally constructed by approximating the profile of the vocal tract using cylindrical tube sections. The resulting tube system is then modeled by a digital ladder filter. This approach was first used in [1] and is called the Kelly–Lochbaum (KL) model.

There are two severe limitations in this basic scheme: 1) the total length of the vocal tract is quantized according to the sampling interval T and thus it cannot be changed smoothly and 2) the length of the tube sections in the model has to be the same. The first restriction prevents exact matching to a given speech spectrum since not all formant frequencies are possible to achieve.

Formerly, two techniques have been proposed to overcome the first problem. Strube [2] proposed that one of the unit delays of the lattice filter could be replaced by a fractional delay element, i.e., a digital filter that approximates a delay smaller than a unit delay. Later the same technique was independently examined by Laine [3].

The other method for the accurate adjustment of the vocal tract length is due to Wu *et al.* [4]. They keep the number of sections in the model constant, but scale the length of each section by a common factor, which may in principle be an arbitrary real number. This is achieved by changing the sampling rate of the model. This, however, does not require that the sample rate of the computer hardware and DA converters should be variable, but the operation is done by software instead. A time-varying FIR interpolating filter is used for sampling rate conversion. A more efficient form of this technique has been reported by Wright and Owens [5].

A drawback of both the techniques described above is that the second restriction is not solved, i.e., the length of each tube section cannot be independently controlled in the discrete-time model. In Strube's method, only one section is continuously controllable while the change of the sampling rate scales the length of all tube sections by the same factor.

This restriction mainly gives rise to difficulties in the control of the model. Namely, the mapping of the physical parameters to the model parameters (reflection coefficients in the scattering junctions) is not simple. The articulators move in the front-back dimension and the diameter of the tract also varies whereas in the model only the diameter of the tube can be adjusted at the fixed points. Control would be more intuitive if the junctions could be moved.

Recently, we have proposed a method for changing the position of the scattering junctions of the KL model continuously [6], [7]. Consequently, the length of every tube section can be continuously varied. This model employs interpolation together with a new technique that we call *deinterpolation*.

In this paper we propose a further improvement to the KL model. Instead of cylindrical tube sections we use *truncated cones*. The resulting model is computationally only slightly more expensive than the basic KL model. Namely, the reflection coefficient at the joint of two conical tubes of different taper is not a real number as in the case of cylindrical tubes, but a first-order filter. The conical-tube approximation of the tract profile is, however, better than that obtained with the same number of cylindrical tube sections. This is quickly understood noticing that conical tubes perform a *first-order polynomial approximation* of the vocal tract profile as opposed to the zeroth-order approximation performed by cylindrical tubes.

The main contribution of the present work is the introduction of digital filters and structures for modeling conical tube systems. Furthermore, we combine the fractional delay junctions and conical tube sections. This results in a novel vocal tract model which is easily controllable and more accurate than the traditional KL model and thus is particularly well suited for articulatory speech synthesis.

II. CONICAL TUBE MODEL

In this section we discuss simulation of conical tubes, and the differences between the KL model and the new one presented in this paper. The basic ideas of discrete-time simulation of non-cylindrical acoustic tubes have been introduced by Smith [8].

Wave Propagation in a Conical Tube

The linear wave equation for a spherical pressure wave traveling in a conical tube is given by

$$\frac{\partial^2 \psi(r,t)}{\partial t^2} = c^2 \frac{\partial^2 \psi(r,t)}{\partial r^2} \tag{1}$$

where c is the speed of sound, r is the distance from the tip of the cone and $\psi(r,t)$ is defined by

$$\psi(r,t) = rp(r,t) \tag{2}$$

where p denotes the acoustic pressure. The well-known traveling wave solution for Eq. (1) is of the form

$$p(r,t) = \frac{f(t-\frac{r}{c})}{r} + \frac{g(t+\frac{r}{c})}{r} \tag{3}$$

where $f(t-\frac{r}{c})$ and $g(t+\frac{r}{c})$ are the components of the spherical wave that travel in the positive and negative r-direction, respectively.

The traveling wave solution given by Eq. (3) may be sampled to obtain the form suitable for digital simulation [8]. Thus we can define

$$p^+(n) = f(nT - \tfrac{r}{c}) \text{ and } p^-(n) = g(nT + \tfrac{r}{c}) \tag{4}$$

The total pressure $p(n)$ at point r is then obtained as

$$p(n,r) = \frac{1}{r}\left[p^+(n) + p^-(n)\right] \tag{5}$$

This means that wave propagation in a conical tube can be simulated using a bidirectional delay line (i.e., a *digital waveguide* [9]). The total pressure at any point is obtained by summing the value of the two components at that point and dividing by the distance from the tip of the cone.

Junction of Conical Sections

In the following we derive equations that govern the reflection and transmission of pressure waves at the junction of two conical tubes of different diameter and taper.

The following conditions must be fulfilled at the junction:

$$p_a^+ + p_a^- = p_b^+ + p_b^- \tag{6}$$

$$u_a^+ - u_a^- = u_b^+ - u_b^- \tag{7}$$

The admittances in different directions are written as

$$Y_a^\pm = \frac{u_a^\pm}{p_a^\pm} \text{ and } Y_b^\pm = \frac{u_b^\pm}{p_b^\pm} \tag{8}$$

The physical interpretation of the above equations is depicted in Fig. 1. Based on Eqs. (7) and (8) we may write

$$p_a^+ Y_a^+ - p_a^- Y_a^- = p_b^+ Y_b^+ - p_b^- Y_b^- \tag{9}$$

Eliminating p_b^+ using Eq. (6) and solving for p_a^- yields

$$p_a^- = \frac{Y_a^+ - Y_b^+}{Y_a^- + Y_b^+} p_a^+ + \frac{Y_b^+ + Y_b^-}{Y_a^- + Y_b^-} p_b^- \tag{10}$$

Thus we obtain equations for the reflection function R^+ and the transmission function $T^- = 1 + R^-$:

$$R^+ = \frac{Y_a^+ - Y_b^+}{Y_a^- + Y_b^+} \text{ and } T^- = \frac{Y_b^+ + Y_b^-}{Y_a^- + Y_b^+} \tag{11}$$

Similarly, we may eliminate p_a^- from Eq. (9) and solve for p_b^+. We then obtain

$$R^- = \frac{Y_b^- - Y_a^-}{Y_b^+ + Y_a^-} \text{ and } T^+ = \frac{Y_a^+ + Y_a^-}{Y_b^+ + Y_a^-} \tag{12}$$

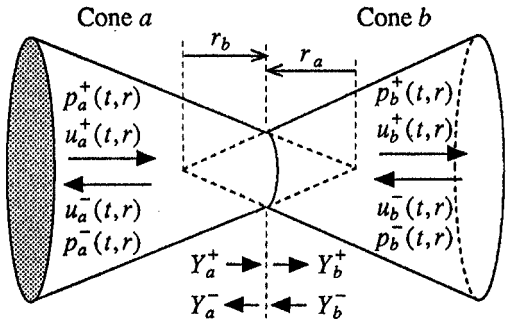

Fig. 1 *A junction of conical tubes that have the same diameter A at the joint. In this example $r_a < 0$ and $r_b > 0$.*

We may now use the expressions for the acoustic admittances of spherical waves in a conical tube

$$Y_a^\pm = \frac{A_a}{\rho c}\left(1 \mp \frac{j}{kr_a}\right) \text{ and } Y_b^\pm = \frac{A_b}{\rho c}\left(1 \mp \frac{j}{kr_b}\right) \tag{13}$$

where $k = \omega / c$. We can derive a general expression for, e.g., the reflection function R^+. This yields

$$R^+(\omega) = \frac{B-1}{B+1} - \frac{2B\alpha}{(B+1)(j\omega + \alpha)} \tag{14}$$

where B is the area ratio of the sections a and b at the joint, i.e.,

$$B = A_a / A_b \tag{15}$$

and α is defined by

$$\alpha = -\frac{c}{A_a + A_b}\left(\frac{A_a}{r_a} - \frac{A_b}{r_b}\right) \tag{16}$$

A major difference between the junction of cylindrical tubes used in the KL model and of conical tubes presented above is that now the scattering is described by means of a filter instead of a real coefficient. Note that in limit $r_a, r_b \to \infty$ the reflection function given by Eq. (14) approaches the reflection coefficient for a junction of two cylindrical tubes.

The expression for the reflection function of a junction of conical tubes of different diameter and taper has also been derived in [10] but using another approach.

Junction with Taper Discontinuity Only

A reflection function needed in a model that approximates the vocal tract profile using linear interpolation is obtained by setting areas A_a and A_b to be equal, i.e., $B = 1$. Substituting this into Eq. (14) we obtain the reflection function for a junction with taper discontinuity only (see Fig. 1), that is

$$R^+(\omega) = -\frac{\alpha}{\alpha + j\omega} \tag{17}$$

where the coefficient α is now given by

$$\alpha = -c / 2r_a + c / 2r_b \tag{18}$$

where r_a and r_b are the distances from the (imagined) tips of cones a and b, respectively, as illustrated in Fig. 1. This reflection function can be converted into a digital filter using the *impulse-invariant transform.* This yields

$$R^+(z) = \frac{b_0}{1 + a_1 z^{-1}} \tag{19}$$

with the filter coefficient $a_1 = -e^{-\alpha T}$ where α is defined as in Eq. (18). It is necessary to scale this filter to have unity gain at $\omega = 0$. This is achieved by setting $b_0 = -(1 + a_1)$.

The impulse-invariant transformation causes aliasing to the digital filter. In this case the effect of aliasing is remarkable only at high frequencies (for examples, see [11]). At low frequencies, the frequency response of the analog reflection filter and that of the digital one coincide.

Above we have considered reflection from the positive side of a junction. In the case $B = 1$, the reflection function $R^-(z)$ from the negative direction is the same as the one from the positive direction.

It appears that in many practical situations the reflection filter $R^+(z)$ defined by Eq. (19) is unstable. Having an unstable filter as a part of a larger system does not necessarily imply that the overall system is unstable. It appears that the digital model of physically realizable conical tube configurations results in a stable system.

Closed End of a Conical Tube

We shall next discuss the boundary conditions of the conical tube model in the glottis. When the end of a truncated conical tube is closed, the radiation admittance tends to zero. Then the reflection function is written as

$$\begin{aligned} R^+(\omega) = \frac{Y_a^+}{Y_a^-} &= \frac{1 - j/kr_a}{1 + j/kr_a} = \frac{j\omega + c/r_a}{j\omega - c/r_a} \\ &= 1 + \frac{2c/r_a}{j\omega - c/r_a} \end{aligned} \tag{20}$$

The impulse-invariant transform of the last form yields

$$R(z) = 1 + \frac{b_0}{1 + a_1 z^{-1}} \tag{21}$$

with the coefficients $b_0 = 2cT/r_a$ and $a_1 = -e^{cT/r_a}$. Note that in the limit $r_a \to \infty$ this transfer function approaches unity, which is the reflection coefficient for the closed end of a lossless cylindrical tube. A more accurate approximation for the reflection at the glottis end may be obtained by substituting a small real value (i.e., a large resistive load) for the radiation admittance instead of zero.

Open End of a Conical Tube

The acoustic impedance of an open end of a conical tube cannot be expressed in closed form. In [10] an *ad hoc* approximation for the reflection function of an open end of a conical tube in an infinite baffle was proposed. This function can also be used for modeling the reflection from the lips in a vocal tract model:

$$r_c(t) = \begin{cases} -a^2 t e^{-at} & t \geq 0 \\ 0 & t < 0 \end{cases} \tag{22}$$

where a is defined by

$$a = c\left(\frac{0.787e}{D} - \frac{1}{2r_e}\right) \tag{23}$$

where D is the diameter of the open end, r_e the distance of the opening from the apex of the cone, and $e \approx 2.71828$.

Equation (22) is easily transformed into the frequency domain employing the well-known property that multiplication by t in the time domain corresponds to differentiation with respect to the frequency variable in the frequency domain. Thus, the Laplace transform of Eq. (22) is

$$R_c(s) = -a^2 \frac{d}{ds}\left(\frac{1}{s+a}\right) = -\frac{a^3}{(s+a)^2} \tag{24}$$

The corresponding z-transform obtained by using the impulse-invariant method is given by

$$R(z) = -a^2 \frac{d}{dz}\left(\frac{1}{1 - e^{-aT} z^{-1}}\right) \tag{25}$$

This yields

$$R(z) = \frac{b_1 z^{-1}}{1 + a_1 z^{-1} + a_2 z^{-2}} \tag{26}$$

with the coefficients

$$b_1 = -a^2 T^2 e^{-aT},\ a_1 = -2e^{-aT},\ a_2 = e^{-2aT} \tag{27}$$

This reflection function is not a very accurate approximation as may be observed by comparing the impulse responses shown in [10]. Another model for the lip reflection may be derived using some well-known radiation admittance as Y_b^+ in R^+ in Eq. (11).

III. FRACTIONAL DELAY TUBE MODEL

The length of each tube section in the vocal tract model can be controlled independently when the propagation delay between the scattering junctions is implemented using fractional delays (FDs) [12], [6], [7]. In the following we discuss implementation of a scattering junction that is not located at an integral point of the delay lines.

In Fig. 2, the block labeled z^{-d} represents an ideal fractional delay of d ($0 \leq d < 1$). The corresponding frequency response is obtained by setting $z = e^{j\omega}$. This yields

$$H(e^{j\omega}) = e^{-j\omega d} \tag{28}$$

Here the sampling rate has been normalized to 1.

The variable δ is called the *complementary fractional delay* and it is defined by

$$\delta = 1 - d \tag{29}$$

Together the fractional delay elements z^{-d} and $z^{-\delta}$ construct a unit delay.

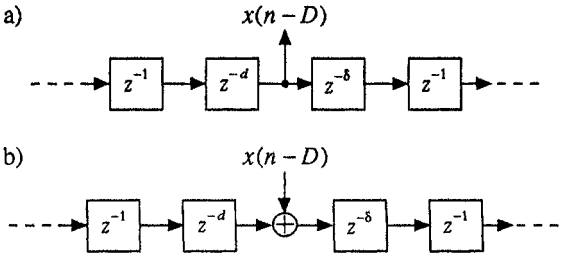

Fig. 2 *a) Output and b) input at a fractional point of a digital delay line.*

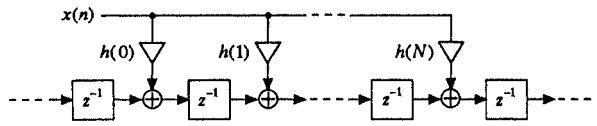

Fig. 3 *The operation of deinterpolation can be implemented using the transpose FIR filter structure.*

Deinterpolation

The system of Fig. 2a requires the use of bandlimited interpolation. The operation used in Fig. 2b to add a signal "between the samples of a delay line" is an inverse operation of interpolation. This operation has been given the name *deinterpolation* [12].

The major difference between interpolation and the corresponding deinterpolation is that the former realizes a fractional delay d while the latter the complementary delay δ. This is seen by considering Fig. 2 again: in interpolation the outgoing signal $x(n-D)$ travels through the element z^{-d} (see Fig. 2a) and in deinterpolation the ingoing signal $x(n-D)$ goes through $z^{-\delta}$ (see Fig. 2b).

Interpolation is usually implemented by the standard direct-form FIR filter structure. Deinterpolation can be implemented by the *transpose FIR filter structure* as illustrated in Fig. 3. Note that the coefficients $h(n)$ of the deinterpolator are in *time-reversed order* with respect to the transpose FIR filter structure since the deinterpolator approximates the complementary delay $z^{-\delta}$.

Interpolated Junction of Conical Sections

The FD scattering junction is constructed by replacing each scattering junction with an interpolated one that is depicted in Fig. 4. The input of the reflection function $R(z)$ is obtained as a sum of the interpolated signal values from both delay lines. Its output is superimposed onto the delay lines using deinterpolation.

A complete vocal tract consists of an arbitrary number of interpolated scattering junctions. The location of each junction can be adjusted by computing the proper coefficients in the FIR interpolation and deinterpolation filters. A time-varying conical tube model can be implemented by changing the filter coefficients as a function of time. The FIR FD filters do not cause disturbing transients if the changes of the interpolation point are not large.

IV. CONCLUSION

A method for discrete-time modeling of the vocal tract using conical tube sections was developed. In addition, it was shown how the formerly introduced fractional delay filtering techniques can be incorporated in this new model.

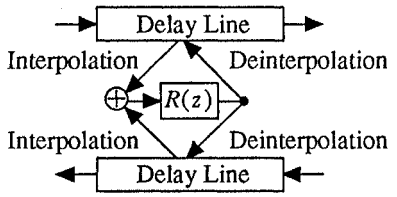

Fig. 4 *The interpolated scattering junction. The input signal of $R(z)$ is obtained as a sum of signals that are interpolated from the delay lines. The output signal is superimposed onto both delay lines using deinterpolation.*

This paper discussed the theoretical aspects of the conical tube simulation. Since the resulting vocal tract model is believed to be more accurate and intuitively controllable than the traditional KL model, we are planning to use it for articulatory speech synthesis.

ACKNOWLEDGMENTS

The authors are grateful to Dr. Paavo Alku for his comments on an earlier version of this paper. This study has been financially supported by the Academy of Finland.

REFERENCES

[1] J. L. Kelly Jr. and C. C. Lochbaum, "Speech synthesis," in *Proc. Fourth Int. Congr. Acoust.*, Paper G42, pp. 1–4, Copenhagen, 1962.

[2] H. W. Strube, "Sampled-data representation of a non-uniform lossless tube of continuously variable length," *J. Acoust. Soc. Am.*, vol. 57, no. 1, pp. 256-257, Jan. 1975.

[3] U. K. Laine, "Digital modelling of a variable-length acoustic tube," in *Proc. Nordic Acoustical Meeting*, Tampere, Finland, June 15–17, 1988, pp. 165–168.

[4] H. Y. Wu, P. Badin, Y. M. Cheng, and B. Guerin, "Continuous variation of the vocal tract length in a Kelly–Lochbaum type speech production model," in *Proc. Int. Congr. Phonetic Sciences (XIth ICPhS)*, Tallinn, Estonia, August 1–7, 1987, pp. 340–343.

[5] G. T. H. Wright and F. J. Owens, "An optimized multirate sampling technique for the dynamic variation of the vocal tract length in the Kelly-Lochbaum speech synthesis model," *IEEE Trans. Speech and Audio Processing*, vol. 1, no. 1, pp. 109-113, Jan. 1993.

[6] V. Välimäki, M. Karjalainen, and T. Kuisma, "Articulatory speech synthesis based on fractional delay waveguide filters," in *Proc. 1994 IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 1, pp. 571–574, Adelaide, Australia, April 19–22, 1994.

[7] V. Välimäki, M. Karjalainen, and T. Kuisma, "Articulatory control of a vocal tract model based on fractional delay waveguide filters," in *Proc. 1994 Int. Symp. Speech, Image Processing and Neural Networks (ISSIPNN'94)*, vol. 2, pp. 571–574, Hong Kong, April 13–16, 1994.

[8] J. O. Smith, "Waveguide simulation of non-cylindrical acoustic tubes," in *Proc. Int. Computer Music Conf.*, Montreal, pp. 304–307, Oct. 16–20, 1991.

[9] J. O. Smith, "Physical modeling using digital waveguides," *Computer Music J.*, vol. 16, no. 4, pp. 75–87, winter 1992.

[10] J. Martínez and J. Agulló, "Conical bores. Part I: reflection functions associated with discontinuities," *J. Acoust. Soc. Am.*, vol. 84, no. 5, pp. 1613–1619, Nov. 1988.

[11] V. Välimäki and M. Karjalainen, "Digital waveguide modeling of wind instrument bores constructed of truncated cones," in *Proc. 1994 Int. Computer Music Conf.*, Aarhus, Denmark, to be published.

[12] V. Välimäki, M. Karjalainen, and T. I. Laakso, "Fractional delay digital filters," in *Proc. 1993 IEEE Int. Symp. Circuits and Systems (ISCAS'93)*, Chicago, IL, vol. 1, pp. 355–358, May 3–6, 1993.