IEEE ICASSP-94, April 19-22, 1994, Adeleide, Australia

ARTICULATORY SPEECH SYNTHESIS BASED ON FRACTIONAL DELAY WAVEGUIDE FILTERS

Vesa Välimäki, Matti Karjalainen, and *Timo Kuisma*

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

ABSTRACT

An extension to the traditional Kelly-Lochbaum vocal tract model is introduced. In the new model not only the diameter but also the length of each tube section can be continuously adjusted. This is achieved by using fractional delay filter techniques such as interpolation and deinterpolation. The filter structure consisting of bidirectional delay lines (digital waveguides) and interpolated ports that connect two or more waveguide sections together is called a fractional delay waveguide filter (FDWF). The interpolated version of the two-port scattering junction is presented and a technique for analyzing the degradation due to approximation errors in interpolation and deinterpolation is described. It is shown that when an FDWF structure with Lagrange interpolation is used a vocal tract model needs to be implemented using oversampling. For example, a sampling rate of 22 kHz is adequate for producing high-quality synthetic sounds at a 5 kHz bandwidth.

1. INTRODUCTION

Speech synthesis models are traditionally divided into two main categories: line analogs and terminal analogs. The former are well suited to articulatory synthesis where the time-varying cross-sectional area of the vocal tract is approximated by an equivalent electronic or computational transmission line. The latter simulate the transfer function from the excitation to the radiated sound by a controllable filter system (e.g. cascaded or parallel resonators).

The line analog models approximate the shape of the vocal tract by a cascade of tube sections and junctions so that each section is a uniform (often lossless) transmission line. This is the so called Kelly-Lochbaum (KL) model [1]. It is convenient and natural for digital implementation since the propagation delays between the connections of tubes and scattering junctions can be computed very efficiently. For a comprehensive analysis of this technique for articulatory speech synthesis, see [2].

A fundamental limitation of the traditional KL model is that the length of the tube sections and the whole system always corresponds to an integer multiple of the sample interval. There have been attempts to adjust the total delay length of the vocal tract model more accurately, e.g., by various interpolation techniques [3]–[5] or by using polynomial approximation of a fractional delay [6]. However, these works do not suggest any systematic method for changing the length of individual subsections.

In this paper we show how it is possible to continuously adjust the positions of KL junctions as well as the total length of the vocal tract by means of *fractional delay waveguide filters* (FDWF). They were introduced recently by Välimäki *et al.* [7] as a generalization of *digital waveguide filters* [8] that are suitable elements for the simulation of one-dimensional waveguides. By the term FDWF we mean a discrete-time structure that consists of bidirectional delay lines (digital waveguides) and fractional delay ports that connect the waveguide sections together. This formalism provides a natural and relatively efficient way for spatially continuous discrete-time simulation of one-dimensional wave propagation and scattering. In addition to speech synthesis FDWFs have been applied to music synthesis, e.g., to the physical modeling of woodwind instruments with finger holes [9].

The paper is organized as follows. In Section 2 we describe the basic ideas for an FDWF-based model for the vocal tract. Section 3 discusses the interpolated KL junction. A method for error analysis of an FD junction and simulation results are presented in Section 4 while Section 5 includes a short discussion on implementation issues. Finally, Section 6 concludes with some hints on our future work.

2. SPEECH SYNTHESIS MODEL WITH VARIABLE-LENGTH SECTIONS

The applicability of variable-length tube sections to the modeling of speech production was shown in the classical works of Fant [10] for example. The idea is well suited to the principle of moving articulators. Especially the tongue forms an independent section that moves in the front-back and up-down dimension controlling the cross-sectional area of the constriction. The opening of the lips and the entire length of the vocal tract are also variable. Even the larynx moves in the up-down direction. Thus it seems natural to divide the tract into a small number of continuously controllable sections that follow the parameters of the articulatory organs in an inherent way.

Figure 1 illustrates a three-section model of the vocal tract. The delays τ_m, that correspond to the physical length of the uniform tube sections, are controlled in a continuous way. Thus also the total length of the system is continuously variable. The junctions k_m represent a discontinuity of the acoustic impedance, i.e., change of the cross-sectional area of the tube. The related wave scattering may be implemented as a KL two-port [1]. The reflection coefficients r_m associated with the ports are computed as

$$r_m = \frac{Z_{m+1} - Z_m}{Z_{m+1} + Z_m} = \frac{A_m - A_{m+1}}{A_{m+1} + A_m} \tag{1}$$

where Z_m and A_m denote the acoustic impedance and cross-sectional area of the mth tube section, respectively.

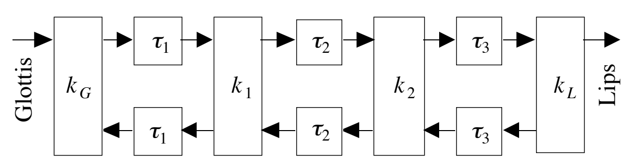

Fig. 1. A three-section vocal tract model. Blocks k_m are KL junctions with different reflection coefficients r_m and blocks τ_m represent the propagation delay between the junctions.

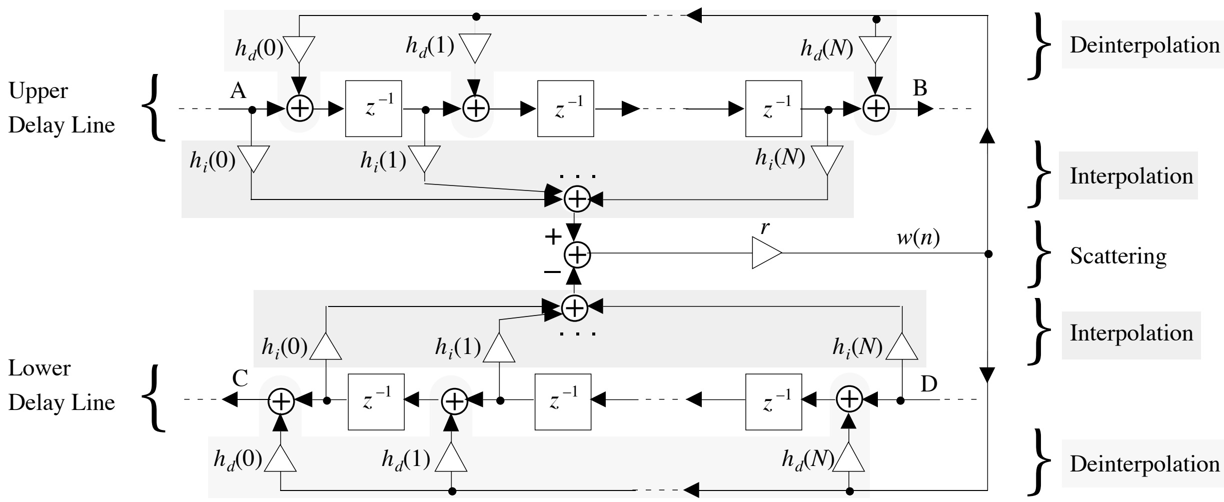

Fig. 2. Implementation of a fractional delay two-port for pressure signals using FIR interpolation and deinterpolation.

The glottis may be implemented as a rapidly opening and closing special section. The lip radiation may be realized using a first-order IIR filter as proposed in [11]. The nasal tract could be connected as a side branch via a three-port [7], [9].

3. THE INTERPOLATED KL JUNCTION

An FDWF-based vocal tract model is composed of two delay lines (a digital waveguide) and a number of interpolated ports. Figure 2 illustrates the flow diagram of a single fractional delay two-port for pressure signals. The wave scattering component is computed by multiplying the difference of the interpolated signals with the reflection coefficient r. Below we describe the fractional delay KL junction that by far is the most important element in our speech synthesis model.

3.1. FIR Interpolation for Fractional Delay Approximation

The signals traveling in the upper and lower lines (see Fig. 2) are interpolated out from the lines by FIR-type interpolating filters with the transfer function

$$H_i(z) = \sum_{n=0}^{N} h_i(n) z^{-n} \tag{2}$$

where N is the order of the FIR filter, and the coefficients $h_i(n)$ approximate the impulse response of the ideal interpolator. An especially suitable method for bandlimited interpolation using an FIR filter is Lagrange interpolation which is a maximally flat approximation (at $\omega = 0$) of the ideal fractional delay [12], [13]. The coefficients of a Lagrange interpolator are [3]

$$h_i(n) = \prod_{k=0, k\neq n}^{N} \frac{D-k}{n-k} \quad \text{for } n = 0,1,\ldots,N \tag{3}$$

where $D = \text{floor}(D) + d$, $d \in \Re$ is the desired delay in samples. In order to minimize the approximation error D should be chosen so that $(N-1)/2 \le D \le (N+1)/2$.

3.2. FIR Deinterpolation

The scattered signal $w(n)$ is fed back to the delay lines by *deinterpolation* [7], i.e., superposition of the signal to the lines by using the transpose of the FIR filter (with interpolating coefficients). When the coefficients $h_i(n)$ are used also in deinterpolation, the transfer function $H_d(z)$ of the deinterpolation filter is equal to that of the interpolation filter, i.e., $H_d(z) = H_i(z)$. The output of the deinterpolating filter is computed as

$$\hat{s}(k-n) = s(k-n) + h_d(n) x(k-D) \quad \text{for } n = 0,1,\ldots,N \tag{4}$$

where $h_d(n)$ are the coefficients of the deinterpolation filter, $x(k-D)$ is the signal sample to be deinterpolated to the point D, and $s(k)$ and $\hat{s}(k)$ are the signals in the delay line before and after deinterpolation, respectively.

3.3. Transfer Functions through the FD Two-Port

In order to get intuition to the functioning of the FD port depicted in Fig. 2, we consider the transmission and reflection functions of the interpolated two-port from both sides. In the following we assume that $h_i(n) = h_d(n)$ for all n. The transfer function from A to B (in Fig. 2) can then be written as

$$\begin{aligned} T^+(z) &= z^{-N}\left[1 + rH_i(z)H_d(z^{-1})\right] \\ &= z^{-N} + r\sum_{n=0}^{N} h_i(n) \sum_{m=0}^{N}\left[h_d(N-m) z^{-n-m}\right] \end{aligned} \tag{5}$$

It can be seen from this z-transform that the impulse response (IR) through the port consists of a delayed unit impulse plus the convolution of $h_i(n)$ and the time-reversed (and delayed) IR $h_d(N-n)$ scaled by r. The length of the IR is thus $2N+1$. Since $H_d(z) = H_i(z)$, the phases of these transfer functions cancel each other out and a *linear-phase IR* results. A transfer function corresponding to a linear-phase filter is also obtained from D to C (see Fig. 2)

$$T^-(z) = z^{-N}\left[1 - rH_i(z^{-1})H_d(z)\right] \tag{6}$$

The reflection function of the FD two-port from A to C in Fig. 2 is

$$R^+(z) = rH_i(z)H_d(z) = r\sum_{n=0}^{N} h_i(n) \sum_{m=0}^{N}\left[h_d(m) z^{-n-m}\right] \tag{7}$$

Again we can see that a finite IR of length $2N+1$ is obtained. This IR is a convolution of the IRs of the interpolation and deinterpolation filter (scaled by r), and does not have linear phase unless both $H_i(z)$ and $H_d(z)$ are linear-phase transfer

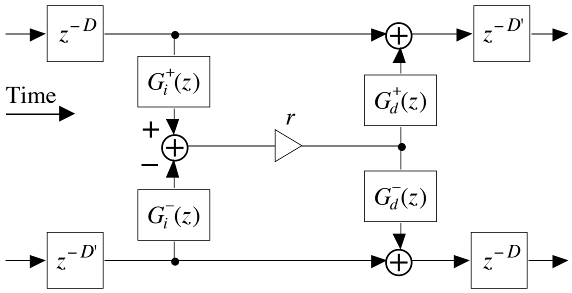

Fig. 3. Temporal flow diagram with ideal fractional delays D and D' The transfer functions $G_i^+(z)$ and $G_i^-(z)$ include the interpolation error and $G_d^+(z)$ and $G_d^-(z)$ the deinterpolation error, and r is the reflection coefficient.

functions. The reflection function from D to B (see Fig. 2) can be written as

$$R^-(z) = -rz^{-2N} H_i(z^{-1}) H_d(z^{-1}) \tag{8}$$

Here the transfer functions of both the interpolation and deinterpolation filters appear in the time-reversed and delayed form.

4. ANALYSIS OF APPROXIMATION ERRORS

The approximation errors of the fractional delay filters degrade the performance of the speech synthesis model. The interpretation of Fig. 3 helps in the error analysis. The block z^{-D} denotes the transfer function of the ideal fractional delay corresponding to the Fourier transform $e^{-j\omega D}$ where $\omega = 2\pi fT$ is the normalized angular frequency with sample interval T. The delay D' in Fig. 3 is defined as

$$D' = N - D \tag{9}$$

The transfer functions $G(z)$ represent the difference of the ideal signal path and the approximation error, i.e.,

$$\begin{aligned} G_i^+(z) &= z^D H_i(z) = 1 - z^D E_i^+(z), \\ G_i^-(z) &= z^{D'} H_i(z^{-1}) = 1 - z^{D'} E_i^-(z), \\ G_d^+(z) &= z^{D'} H_d(z^{-1}) = 1 - z^{D'} E_d^+(z), \\ G_d^-(z) &= z^D H_d(z) = 1 - z^D E_d^-(z) \end{aligned} \tag{10}$$

where the error transfer functions are given by

$$\begin{aligned} E_i^+(z) &= z^{-D} - H_i(z), \quad E_i^-(z) = z^{-D'} - H_i(z^{-1}), \\ E_d^+(z) &= z^{-D'} - H_d(z^{-1}), \quad E_d^-(z) = z^{-D} - H_d(z). \end{aligned} \tag{11}$$

4.1. Magnitude Errors in an FDWF-Based Tube Model

The approximation error of Lagrange interpolation is zero at $\omega = 0$ by definition of the maximally flat filter design [13]. The magnitude response is flat at all frequencies only when D is an integer, i.e., when the interpolator is reduced to a trivial integer delay. Otherwise, the Lagrange interpolator is a low-pass filter and the error transfer functions in Eq. (11) are thus high-pass filters. In the worst case, $d = 0.5$, the Lagrange interpolator has a zero at the Nyquist frequency. The phase error appears as a nonlinearity of the phase at high frequencies except when $d = 0.5$, since then the interpolator is a linear-phase filter.

We have computed the magnitude response of a two-tube model with Lagrange interpolation and with ideal interpolation

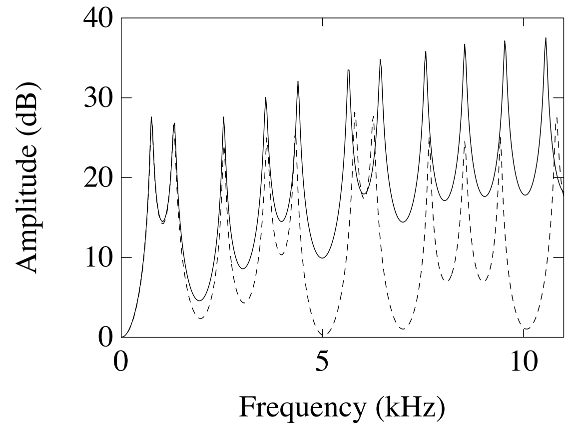

Fig. 4. Comparison of the worst-case ($d = 0.5$) magnitude response of a two-tube model using linear interpolation and deinterpolation (solid line) with the ideal spectrum (dashed line). The sampling rate in this example is 22 kHz.

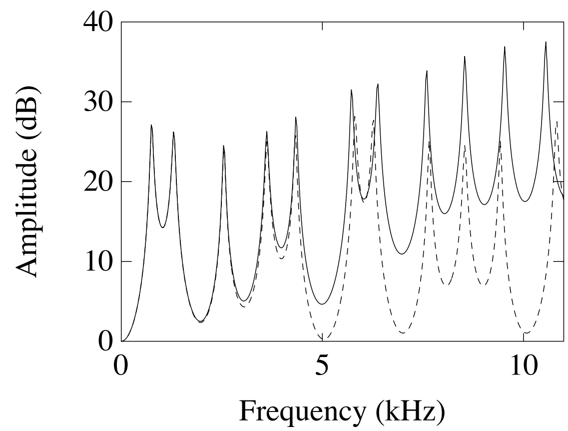

Fig. 5. The magnitude responses of a two-tube model with third-order Lagrange interpolation and deinterpolation in the worst case (solid line) and with ideal fractional delays (dashed line).

using a sampling rate of 22 kHz. The solid line in Fig. 4 is obtained using first-order Lagrange interpolation (linear interpolation) with $d = 0.5$. The dashed line is the ideal magnitude response obtained by setting the errors $E(z)$ to zero. Note that at low frequencies the two curves join together. At high frequencies the interpolation error causes the spectral level and the frequencies of the formants to approach those of the neutral vocal tract. This can be understood by the fact that the reflection coefficient is effectively multiplied by the approximation errors (see Fig. 3) and thus reflection becomes inefficient at high frequencies. This implies that as the frequency increases, the frequency response of the two-tube model approaches that of a single uniform tube of the same length.

Figure 5 illustrates the magnitude response of the two-tube model with third-order Lagrange interpolation and with ideal fractional delays. Now the accuracy is relatively good (within 1 dB) up to almost 5 kHz. Above this frequency modes other than the plane wave can also propagate in a vocal tract, and thus even the ideal tube model—which assumes propagation of plane waves only—is not valid anymore. This bandwidth is also wide enough for high-quality speech sounds except for some fricatives. Hence, good choices for parameters of an FDWF-

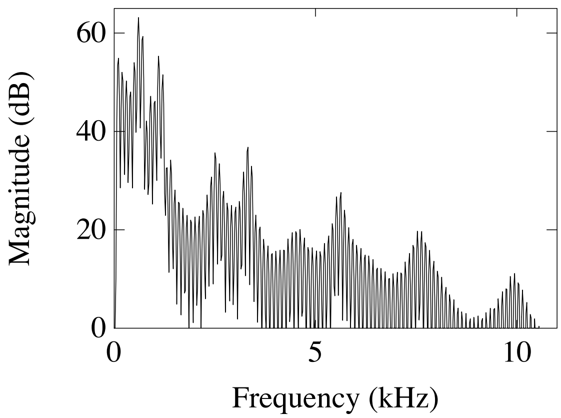

Fig. 6. The magnitude spectrum of a synthetic vowel /*a*/ produced by a four-tube vocal tract model. Third-order Lagrange interpolation and deinterpolation were employed. The sampling rate was 22 kHz.

based speech synthesizer are, e.g., N = 3 and f_s = 22 kHz. The emphasis of the transfer function at high frequencies does not matter much if the glottal excitation is a low-pass signal.

Adding a new FD port to the system increases the approximation error. Then the frequency response will approach that of a uniform tube more quickly as the frequency increases. However, according to our experiments, good results can be obtained with several FD junctions when a high enough sampling rate and high-order fractional delay filters are used. Figure 6 shows an example of the spectrum of the vowel /*a*/ synthesized by a four-tube FDWF model.

5. IMPLEMENTATION ISSUES

The implementation of the interpolated KL junction illustrated in Fig. 2 is not the most efficient one, but the most straightforward. In practice, the two interpolators can be combined into one that interpolates the difference of the sample values in the delay lines. Also, the deinterpolation filters can be joined together so that copies of $w(n)$ are multiplied by each of the N+1 deinterpolating coefficients and the same results are added to the upper and the lower delay lines. Thus, both interpolation and deinterpolation need to be computed *once per junction* (not twice as it would seem at first). This kind of organization of the computation results in $2N + 3$ multiplications, $2N + 1$ normal additions, and $2N + 2$ additions to the delay line (which are more expensive than plain addition) for each junction.

We have implemented a speech synthesizer following the principles discussed in this paper on a TMS320C30 signal processor. A five-tube model utilizing third-order FIR interpolation and deinterpolation runs in real time with a sampling rate of 22 kHz. We have synthesized vowels and voiced sounds with dynamic transitions and have also carried out some experiments to synthesize nasals. The results have been most promising.

6. CONCLUSION

A modified version of the KL vocal tract model was presented. In this model each tube section can have an arbitrary length since the KL junctions can be located between the unit delays along the bidirectional delay line. Due to approximation errors the FDWF-based synthesis system has performance problems at high frequencies. This, however, is not a major limitation in practice when a high enough sampling frequency is utilized. For example a sampling rate of 22 kHz with third-order Lagrange interpolation guarantees good performance up to about 5 kHz which is adequate for producing high-quality synthetic voices.

We are developing control strategies for the FDWF-based articulatory synthesizer in order to construct a phoneme-to-speech system. Variable-length tube sections help to relate parts of the model to real world objects, such as the tongue or the velum. The results will be reported in a forthcoming paper.

ACKNOWLEDGMENT

The authors would like to thank Dr. Unto K. Laine for helpful discussions. We are also grateful to CARTES (Espoo, Finland) for the possibility to use their facilities. Part of this study has been financed by the Academy of Finland.

REFERENCES

[1] J. L. Kelly Jr. and C. C. Lochbaum, “Speech synthesis,” in *Proc. Fourth ICA*, Paper G42, Copenhagen, 1962.

[2] J. Liljencrants, *Speech Synthesis with a Reflection-Type Line Analog*. Doctoral dissertation, Dept. of Speech Communication and Music Acoustics, Royal Institute of Technology, Stockholm, Sweden, Aug. 1985.

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

[4] 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.

[5] 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. XIth ICPhS*, pp. 340-343, Tallinn, Estonia, Aug. 1987.

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

[7] V. Välimäki, M. Karjalainen, and T. I. Laakso, “Fractional delay digital filters,” in *Proc. 1993 IEEE Int. Symp. on Circuits and Systems*, pp. 355-358, Chicago, IL, May 1993.

[8] J. O. Smith, “Physical modeling using digital waveguides,” *Computer Music J.*, vol. 16, pp. 75-87, Winter 1992.

[9] V. Välimäki, M. Karjalainen, and T. I. Laakso, “Modeling of woodwind bores with finger holes,” in *Proc. 1993 Int. Computer Music Conf.*, pp. 32-39, Tokyo, Sept. 1993.

[10] G. Fant, *Acoustic Theory of Speech Production*. Mouton, The Hague, 1960.

[11] U. K. Laine, “Modelling of lossy vocal tract in z-domain,” *Report No. 30*, Acoustics Laboratory, Dept. of Electrical Eng., Helsinki Univ. of Technology, Espoo, Finland, 1988.

[12] T. I. Laakso, V. Välimäki, M. Karjalainen, and U. K. Laine, “Real-time implementation techniques for a continuously variable digital delay in modeling musical instruments,” in *Proc. 1992 Int. Computer Music Conf.*, pp. 140-141, San Jose, CA, Oct. 1992.

[13] T. I. Laakso, V. Välimäki, M. Karjalainen, and U. K. Laine, “Fractional delay approximation using digital filters—a tutorial review,” unpublished manuscript, March 1993.