1994 International Symposium on Speech, Image Processing and Neural Networks, 13-16 April 1994, Hong Kong

Articulatory Control of a Vocal Tract Model Based on Fractional Delay Waveguide Filters

Vesa Välimäki, *Matti Karjalainen*, and *Timo Kuisma*

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

Abstract

A novel technique to implement and control an acoustic tube model of the human vocal tract is introduced. This model is an extension to the traditional Kelly–Lochbaum model, since not only the diameter of individual uniform tube sections but also their length, i.e., the positions of scattering junctions, can be continuously varied. The vocal tract model is implemented by means of FIR-type interpolation and deinterpolation that are used to locate the junctions. We show that in this kind of model the articulators can be presented in a natural manner enabling easy control of the model from articulatory data.

1 Introduction

Articulatory models for the vocal tract are typically constructed from a set of uniform tube sections whose length is determined by the sampling rate [1, 2]. For example, when the sampling rate is 16.0 kHz all the sections are 2.2 cm long (with a sound velocity of 350 m/s). Therefore, a 17.5 cm long vocal tract of a man would be modeled by 8 uniform tube sections. This is the traditional Kelly–Lochbaum (KL) speech production model.

There are two fundamental limitations in the KL model: 1) it is not possible to adjust the length of the vocal tract continuously since the total length is a multiple of the length of an individual section, and 2) the diameter of the vocal tract can not be changed at arbitrary locations along the tract but only at the junctions of the tube sections.

The total length of the vocal tract of an adult male varies continuously between 16 and 19.5 cm depending on the phonemes. It should also be possible to have this variation in the model. Two solutions to this problem have been proposed: 1) change the sampling rate of the system to vary the length of the individual tube sections [3, 4] or 2) include a fractional delay filter in the system to fine-tune the length of one of the sections [5, 6]. The problem of adjusting the position of scattering junctions is not solved by these techniques.

In our approach the diameter of the tube sections may change at any desirable points, not only at the sampling points. This is possible because the scattering junctions are located at arbitrary points along the tract by using interpolation techniques as shown in [7] and [8]. An additional advantage of this technique is that the sampling rate of the system need not be changed.

The use of fractional delays in modeling the vocal tract allows for a natural mapping from the articulators and their parameters to the sections of the waveguide model. We show that this makes the control of the model intuitive since now it is more straightforward than before to associate parts of the model to physical reality.

2 Fractional Delay Waveguide Filters and Speech Synthesis

Fractional delay waveguide filters (FDWF) were introduced recently by Välimäki *et al.* [7]. They can be seen as a generalization to digital waveguide filters developed by Julius Smith [9]. Digital waveguides (bidirectional delay lines) are suitable elements for the simulation of one-dimensional resonators, such as acoustic tube sections. An FDWF is composed of digital waveguides and a number of interpolated ports to connect together sections of different impedance. FIR-type interpolation techniques are used to fine-tune the lengths of the delay lines and the positions of the ports.

The two-port scattering junction is the most important element of the KL speech synthesis model. It is used to implement a discontinuity of acoustic impedance and the related wave scattering. Figure 1 illustrates the one-multiplier form of the traditional KL junction for volume velocity waves. The reflection coefficient r_m is defined as

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

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

A three-port can be used to connect a nasal tract model as a side branch to the system. It is needed in the synthesis of nasal sounds, e.g., /n/ or /m/. Earlier we have used a similar port for modeling the tone holes of woodwind instruments [10].

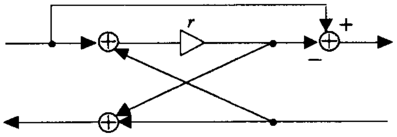

Figure 1: The one-multiplier KL junction.

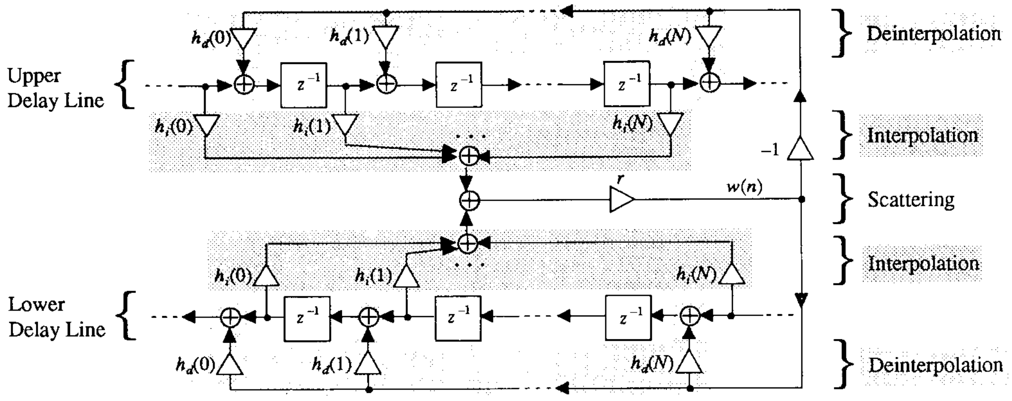

Figure 2: The fractional delay two-port for volume velocity implemented using FIR interpolation and deinterpolation.

2.1 The Interpolated KL Junction

We use the one-multiplier form of the KL junction as the basis for the FD junction which is a two-port that is located at an arbitrary point between two unit delay elements of a delay line. This interpolated KL junction is depicted in Fig. 2. The signals traveling in the upper line to the right and in the lower line to the left are interpolated out from the lines by FIR-type interpolators. These outputs approximate the signals at the position D = floor(D) + d with the fractional delay $d \in \Re$. The model is thus virtually continuous in space although discrete in time. The wave scattering component is computed as the product of the sum of the interpolated wave signals and the reflection coefficient r. This signal is fed back to the delay lines by **deinterpolation** [7, 8], i.e., superposition by means of the transposed FIR interpolation structure (see Fig. 2).

2.2 Interpolation and Deinterpolation Coefficients

Interpolation coefficients $h_i(n)$ and deinterpolation coefficients $h_d(n)$ are the same, e.g., Lagrange interpolation coefficients [8] defined as

$$h_i(n) = \prod_{k=0,k\neq n}^{N} \frac{D-k}{n-k} \quad \text{for } n = 0,1,\ldots,N \qquad (2)$$

where N is the order of the FIR filter and D is the desired delay. In order to minimize the approximation error, D should be chosen so that

$$\frac{N-1}{2} \le D \le \frac{N+1}{2} \qquad (3)$$

In other words, the approximation point should lie between the central taps of the interpolator (in the case of odd N) or within half a sample from the middle tap (with even N). Strube [5] and Laine [6] also used Lagrange interpolation for fractional delay approximation. However, their aim was to adjust the length of the delay line, not directly the locations of the scattering junctions.

3 Articulatory Speech Synthesis Model with Variable-Length Sections

The applicability of variable-length tube sections to the acoustic analysis of speech production was shown in classical works such as [11]. The idea suits well to the principle of moving articulators. Especially the tongue forms a section that moves in the front-back dimension. The cross-sectional area of the constriction is controlled as well and the opening of the lips and the entire length of the vocal tract are 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. Earlier, strategies to map the articulatory control parameters to the vocal tract parameters have been presented [12–15]. In this work, however, we show that a more direct and intuitive model can be constructed when articulatory parameters are used with interpolated KL junctions.

A simple three-section model of the vocal tract is shown in Fig. 3. Each junction k_m may be implemented as an interpolated two-port of Fig. 2 The diagram in Fig. 3 illustrates the possibility to associate certain articulators, like the tongue, to some part of the model. Now the

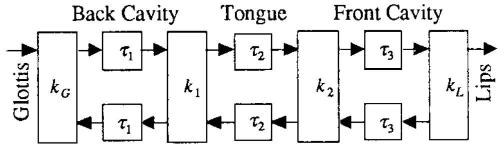

Figure 3: A three-section model for articulatory speech synthesis based on fractional delay waveguide filters. Blocks k_m are KL junctions with different reflection coefficients r_m and blocks τ_m are fractional delays.

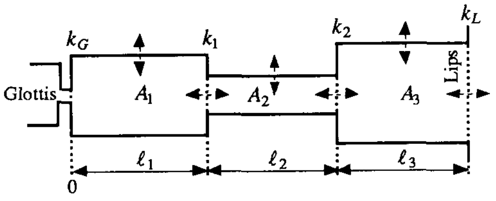

Figure 4: A three-tube model of the vocal tract. Arrows indicate the variable parts. Compare with Fig. 3 above.

movement of the tongue can be simulated by moving the junctions k_1 and k_2 along the bidirectional delay line. This is achieved by computing new values for the interpolation/deinterpolation coefficients according to Eq. (2). The reflection coefficient r_m of the junctions is computed from the cross-sectional areas of the tubes as shown in Eq. (1).

Figure 4 presents the tube model corresponding to Fig. 3 showing the variable parameters. They are the cross-sectional areas A_m of the three tube sections and the lengths ℓ_m of the sections. In the model, however, the lengths of the sections are not control parameters, but the locations of the junctions. The location of junction k_G is not changed, but it is considered as the origin. In this case the junctions k_1, k_2 and k_L are interpolated.

The delays τ_m of the computational model (see Fig. 3) are related to the lengths ℓ_m of individual sections (see Fig. 4) in the following way.

$$\tau_m = \frac{\ell_m}{c} \qquad (4)$$

where c is the sound velocity. By changing the lengths ℓ_m also the total length $\sum \ell_m$ of the system can be varied.

In order to make a complete physical model of speech production, the glottis should be implemented as a rapidly opening and closing section. We have, however, used a linear model where we feed a synthetic glottis pulse to the model. The traditional piston-in-sphere model has been used for the lip radiation. An IIR-type filter has been designed to account for the reflection at the lips [16].

4 Articulatory Control Parameters

In the following we propose a strategy to control the FDWF-based speech synthesizer. There have been several attempts in the literature of speech processing to develop a parameter set that would be complete in the sense that every phoneme in the language could be specified with only those parameters [12, 15].

Figure 5 illustrates the five-tube model that we have used for speech synthesis experiments. The locations p_m (m = 0, 1,..., 4) correspond to junctions k_m and p_L to the location of the lip end k_L. The glottis junction k_G is located at the origin of the coordinate system.

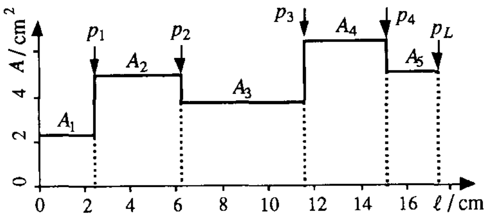

Figure 5: A five-tube model of the vocal tract.

Table 1: The articulatory control parameters and the associated model parameters. See Fig. 5 for model parameters.

Articulatory Parameter	Model Parameter
Jaw opening	all Am
Tongue position	p2 and p3
Dorsal parameter	A3
Tongue tip height	A4
Lip aperture	A5 and pL

4.1 The Parameter Set

The five articulatory parameters used in our vocal tract model are presented in Table 1. Also, the model parameters that have been connected to the articulatory parameters are shown. Note that this model does not include the nasal tract and thus the control of the velum is not included.

The jaw parameter affects all the sections of the vocal tract model. When the jaw is opened, i.e., the value of this parameter is increased, the area values near the lips are increased and those near the glottis are decreased.

The other parameters have more localized effect on the model. The tongue position parameter sets the positions of junctions p_2 and p_3, and the area A_3 between them depends on the dorsal parameter which determines the bulging of the tongue. The height of the tongue tip affects the area A_4, and the lip aperture affects both the area A_5 and p_L. The area A_5 is linearly proportional to the lip aperture parameter whereas p_L is inversely proportional to it.

When producing front vowels, the sections 3 and 4 may be combined. Thus one of the junctions may effectively disappear turning the system temporarily into a four-tube model. Consequently, sections 2 and 3 can be combined when generating back vowels.

For producing nasals a velum parameter would be needed to determine the opening of the velum.

5 Approximation Errors and Transient Behavior

We have synthesized vowels and transitions between them with our synthesizer and have thus gained an understanding of the approximation and dynamic errors of the system. The approximation errors of the FDWF-based model and their analysis are discussed in [8]. It appears that the errors due to Lagrange interpolation take place at high frequencies where the spectrum of the model approaches that of a uniform tube.

According to our test runs, changing the location of the interpolated ports does not cause any transients if the interpolation and deinterpolation coefficients are updated every cycle. This, however, is computationally expensive and we would not like to update the coefficients more often than every 5–10 ms. Then the transients will become annoying, and unfortunately there are not any known analytical method for suppressing them. Formerly, algorithms for suppressing the transients in the dynamic KL model due to the changes of cross-sectional areas have been introduced (see [2] for a clever technique and references to earlier ones).

6 Discussion

The fractional delay waveguide filter model for the vocal tract has been shown to be a promising new technique for articulatory speech synthesis. In addition to the extra degree of freedom due to variable-length tubes, the new model offers a possibility to naturally associate parts of the system to reality. In this paper we have introduced a strategy to map the articulatory parameters to the parameters of the FDWF-based vocal tract model.

The prototype model has been implemented on a TMS320C30 signal processor. The real-time implementation (sample rate 22 kHz) has given us the possibility to test the control of the model interactively. Preliminary synthesis tests have proved that the FDWF-based synthesis model can produce high-quality vowel and nasal sounds and transitions between them. Further experiments are needed in order to learn how to make use of the variable-length tube sections in the synthesis of consonant sounds, like plosives and fricatives.

An important field of future work will be to develop a strategy for suppressing transients due to changes in junction positions when the parameters update rate is less than the sample rate.

Acknowledgments

The authors are grateful to Mr. Toomas Altosaar for his helpful comments on this text. Part of 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. on Acoustics*, Paper G42, Copenhagen, Denmark, 1962.

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

[3] 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. Eleventh Int. Congress of Phonetic Sciences (XIth ICPhS)*, pp. 340–343, Tallinn, Estonia, Aug. 1987.

[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. 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.

[6] U. K. Laine, "Digital modelling of a variable-length acoustic tube," in *Proc. Nordic Acoustical Meeting (NAM'88)*, pp. 165–168, Tampere, Finland, June 1988. Reprinted in [16].

[7] V. Välimäki, M. Karjalainen, and T. I. Laakso, "Fractional delay digital filters," in *Proc. 1993 IEEE Int. Symp. on Circuits and Systems (ISCAS'93)*, pp. 355–358, Chicago, Illinois, March 1993.

[8] 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 Conf. (ICASSP'94)*, Adelaide, Australia, to be published in April 1994.

[9] J. O. Smith, "Physical modeling using digital waveguides," *Computer Music Journal*, vol. 16, no. 4, pp. 74–87, Winter 1992.

[10] V. Välimäki, M. Karjalainen, and T. I. Laakso, "Modeling of woodwind bores with finger holes," in *Proc. 1993 Int. Computer Music Conf. (ICMC'93)*, pp. 32–39, Tokyo, Japan, Sept. 1993.

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

[12] C. H. Coker, "A model of articulatory dynamics and control," *Proc. of the IEEE*, vol. 64, no. 4, pp. 452–460, April 1976.

[13] P. Rudin, T. Baer, and P. Mermelstein, "An articulatory synthesizer for perceptual research," *J. Acoust. Soc. Am.*, vol. 70, no. 2, pp. 321–328, Aug. 1981.

[14] P. Meyer, R. Wilhelms, and H. W. Strube, "A quasiarticulatory speech synthesizer for German language running in real time," *J. Acoust. Soc. Am.*, vol. 86, no. 2, pp. 523–539, Aug. 1989.

[15] B. J. Kröger, "A gestural approach for controlling an articulatory speech synthesizer," in *Proc. 3rd European Conf. on Speech Communication and Tech. (Eurospeech'93)*, pp. 1903–1906, Berlin, Germany, Sept. 1993.

[16] U. K. Laine, *Studies on Modelling of Vocal Tract Acoustics with Applications to Speech Synthesis*. Doctoral thesis. Report no. 32. Helsinki Univ. of Tech., Faculty of Electrical Eng., Acoustics Lab., Espoo, Finland, 1989.