NEW TECHNIQUES OF VOCAL TRACT MODELING FOR ARTICULATORY SPEECH SYNTHESIS

Matti Karjalainen and Vesa Välimäki
Helsinki University of Technology
Laboratory of Acoustics and Audio Signal Processing
Otakaari 5 A, FIN-02150 Espoo
Finland

ABSTRACT

The theory of fractional delay waveguide filters (FDWF) is presented that is proved to solve the fundamental problem of the fixed sample-position lattice found in traditional discrete-time simulation of transmission-line type systems. Thus the approach is well suited to vocal tract models, in particular to articulatory modeling and speech synthesis, where a simple relation between model parameters and articulator positions is desired.

INTRODUCTION

In the days before digital computers the human vocal tract was modeled by analog circuits and physical equivalents of the real system. Transmission-line type models were considered as being continuous, at least in time but generally in space as well. The corresponding analytic mathematical models of speech production, e.g. [1], reflect the same non-discrete character inherent in many macroscopic physical systems. It was natural to simulate the vocal tract by a chain of tube sections (e.g. two and three-tube models) that had relatively direct correspondence to the positions of the articulators.

Computers and digital signal processing changed the picture by providing very accurate and flexible numeric methods to deal with vocal tract modeling [2]. One specific characteristic in digital simulation of continuous-time systems, however, was not flexible at all. Unit delay, the interval between subsequent samples, dictates much of the behavior in discrete-time systems at high frequencies. Signals must be bandlimited and the high-frequency characteristics of analog systems can in a general case be only approximated.

This limitation is particularly prominent in transmission-line models of the vocal tract. When the sampling frequency is fixed, the physical positioning of known sample points in space is fixed as well. Even a detailed adjustment of the vocal tract length to this digital grid has been somewhat difficult and discussed only in relatively few papers. Earlier techniques for this partial solution of the problem are found in [3], [4], and [5].

We may well accept the discrete-time nature (since speech and hearing are bandlimited) but we do not have to accept the discrete-space characteristics in transmission-line modeling. What we need is fractional delays, mathematically equivalent to ideal bandlimited interpolation. This turns out to be possible to approximate and implement using digital filters. Allowing somewhat increased computational costs one can have a whole bunch of discrete-time models that are virtually continuous in space.

We have developed some basic principles and building blocks for modeling transmission-line type systems, such as acoustic tubes of arbitrary and varying shapes, including the human vocal tract [6]–[10]. In this paper we present how to use digital signal processing to build tube sections of varying lengths and Kelly–Lochbaum type tube models with variable junction positions. It will be shown that not only cylindrical but also conical tube sections can be used, thus leading to more natural shape approximations. The only addition that conical tubes introduce to the KL scattering junction is that a simple reflection filter is needed instead of a real-valued reflection coefficient.

The new fractional delay filter structures (we call them Fractional Delay Waveguide Filters, FDWF) are natural candidates when building vocal tract models for articulatory speech synthesis since moving articulators may be associated directly to subsections of the system. The increased computational cost can be compensated by the ever increasing performance of modern signal processors.

GENERALIZED KELLY–LOCHBAUM MODELING

We start by considering a generalized version of the Kelly–Lochbaum (KL) vocal tract model [11]. Figure 1 depicts a one-multiplier KL scattering junction where the traveling wave components are summed, multiplied by the reflection coefficient r, and injected back into the delay lines. This is a traditional formulation and easily implemented as far as the junctions are aligned with the natural sampling positions.

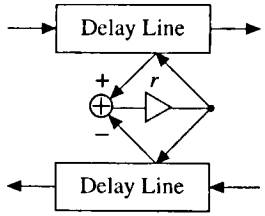

Fig. 1. A one-multiplier Kelly–Lochbaum scattering junction.

Let us consider a case where the scattering junction is located in an arbitrary position along the delay lines. Taking a signal value out of a delay line between sample positions is equivalent to *interpolation*. An ideal bandlimited interpolator is an infinitely long FIR filter with coefficients according to the sinc function [10]. (If the junction is precisely at a sampling point then only that single coefficient remains non-zero.)

The insertion of a signal back into the delay lines, when the insertion point is between the sampling points, is called *deinterpolation* [6]. As an FIR filter it is the transpose of the interpolation filter. The value to be inserted is multiplied by each filter coefficient, these results are added to the sample values in the delay line, and the sums are written back into the sample positions. With ideal interpolators and deinterpolators this precisely implements a bandlimited KL junction in any fractional position.

The two-port KL junction of Fig. 1 can be generalized to a three-port junction that is applicable to the modeling of the nasal tract branch [10].

FRACTIONAL DELAYS

In practice we cannot realize ideal interpolated junctions but have to approximate the infinite series of sinc coefficients by finite order digital filters. Two specific cases are of special interest: (a) Lagrange interpolators of FIR type [3], [5] and (b) allpass filters with maximally flat (at zero frequency) phase delay. Allpass filters have an ideal magnitude response and some other advantages [10]. However, since FIR interpolators are straightforward and conceptually more intuitive, we will consired only them below.

Figure 2 shows the implementation of third-order FIR filters for (2a) an interpolator and (2b) a deinterpolator.

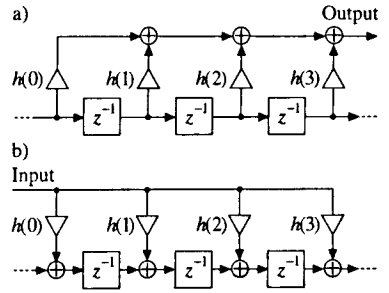

Fig. 2. Third-order FIR filter implementation of (a) an interpolator and (b) a deinterpolator.

When the position of an interpolated junction is moved, the locations of the FIR taps on the delay lines and the values of the filter coefficients must be updated. Notice that there may be any number of fractional delay KL junctions attached on a pair of delay lines. The tap regions of the junctions can also overlap as far as the junction operations are done on a 'read all, compute the scatterings, write all' basis [6].

The filter structure including the delay lines and any number of interpolated

scattering junctions is called the *fractional delay waveguide filter* (FDWF). A special case is a terminating junction where only one interpolation (out) and one deinterpolation (in) is needed. The interpolator of Fig. 2a as such finds many applications where a fractional delay is needed.

The application of finite order interpolators introduces appoximation errors in the waveguide filters. With odd-order Lagrange interpolators the maximum error occurs when the junction is in the middle of two sampling points. An error analysis [10, pp. 93-95] shows that third-order Lagrange interpolators yield good results in speech synthesis up to about 5 kHz when the sampling rate is 22 kHz.

CONICAL TUBE MODEL

The approximation of the vocal tract by cylindrical sections only has been another fundamental limitation of traditional transmission-line models. We have generalized the KL model to allow for conical sections as well (see Fig. 3), since this makes a better match to typical vocal tract shapes that are continuous and relatively smooth functions [9], [10].

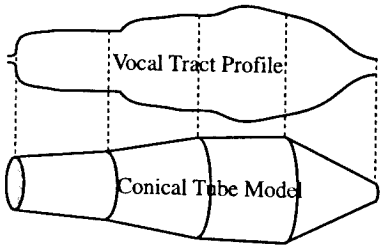

Fig. 3. Conical section approximation of a vocal tract profile.

The difference between cylindrical and conical sections is that instead of plane waves the conical sections carry spherical waves. The derivation of pressure wave scattering at a junction between conical tube sections results in the traditional KL formulation of Fig. 1 except that the reflection coefficient r is replaced by a reflection filter $R(z)$ and the signals from the upper and lower delay line are added (instead of subtraction). The filter $R(z)$ is a first-order IIR filter that is computationally efficient and does not essentially add to the complexity of implementation.

A natural step towards further generality is to combine the freely movable fractional delay ports and the conical sections [9].

ARTICULATORY N-TUBE SYNTHESIS MODEL

Vocal tract modeling using the FDWF techniques is a natural candidate for articulatory speech synthesis [7]. The traditional thinking of the tract as a decomposition of sections related to the articulators [1] is well supported since the section lengths and cross-sectional areas are freely adjustable. Figure 4 shows the case of a three-tube model and the related control parameters.

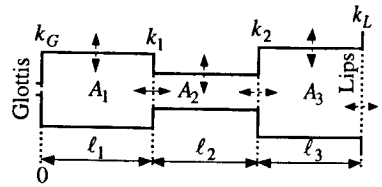

Fig. 4. A three-tube model of the vocal tract. Arrows indicate the variable parts.

The n-tube model parameters have a relatively straightforward mapping from articulatory parameters. In [7] we have proposed a case for a set of five articulatory parameters and a five-tube vocal tract model.

One inherent problem in speech synthesis with models of variable length sections is that there is no simple method known for the analysis of the parameter values from speech signals. Thus iterative or nonlinear estimation techniques must be used.

IMPLEMENTATION ISSUES

We have experimented with the new principles of vocal tract modeling in order to implement real-time synthesis on the TMS320C30 DSP processor.

The computational complexity of the fractional delay waveguide models is relatively high due to the need of interpolation and deinterpolation as well as a relatively high sampling frequency. On the other hand, less tube sections are needed to match a vocal tract profile than with the original KL model.

We have implemented four and five-tube models on the TMS320C30 (33 MHz) floating-point signal processor. Updating of the model parameters has been carried out by a host computer (Apple Macintosh). There is a graphical user interface where the user can move the sections, boundaries, or related articulators by a mouse. We have noticed that the update rate of the parameters should be relatively high (at least every 15 samples when the sampling rate is 22 kHz) in order to avoid audible transient problems. From a theoretical point of view there is need for an analysis of transient-free section length and port position controls in a similar way as was done in [12] for fixed junction models.

So far we have synthesized primarily vowels as well as nasals including a nasal tract. The mouse-controlled vowel synthesizer has been found a useful device for demonstrations and experiments in articulatory phonetics. A full-scale synthesizer with all phoneme classes remains to be developed.

SUMMARY

A non-mathematical introduction to fractional delay waveguide filters has been given and the theory of vocal tract modeling based on them has been presented. The approach allows for a flexible method to implement articulatory speech synthesis using variable-length tube sections. Modeling of the vocal tract by conical tube sections is introduced as another major extension to the theory. A real-time synthesizer with manual tract control has been demonstrated.

REFERENCES

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

[2] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Prentice-Hall, Englewood Cliffs, 1978.

[3] H. W. Strube, "Sampled-data representation of a nonuniform lossless tube of continuously variable length," *JASA*, vol. 57, no. 1, pp. 256–257, Jan. 1975.

[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. XIth ICPhS*, pp. 340–343, Tallinn, Estonia, Aug. 1987.

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

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

[7] V. Välimäki, M. Karjalainen, and T. Kuisma, "Articulatory control of a vocal tract model based on fractional delay waveguide filters," in *Proc. IEEE Int. Symp. Speech, Image Processing and Neural Networks (ISSIPNN'94)*, Hong Kong, vol. 2, pp. 585–588, April 13–16, 1994.

[8] V. Välimäki, M. Karjalainen, and T. Kuisma, "Articulatory speech synthesis based on fractional delay waveguide filters," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP '94)*, Adelaide, Australia, vol. 1, pp. 571–574, April 19–22, 1994.

[9] V. Välimäki and M. Karjalainen, "Improving the Kelly–Lochbaum vocal tract model using conical tube sections and fractional delay filtering techniques," in *Proc. 1994 Int. Conf. Spoken Language Processing (ICSLP'94)*, vol. 2, pp. 615–618, Yokohama, Japan, Sept. 18–22, 1994.

[10] V. Välimäki, *Fractional Delay Waveguide Modeling of Acoustic Tubes*. Report 34, Helsinki Univ. of Tech., Lab. of Acoustics and Audio Signal Processing, Espoo, Finland, 1994.

[11] J. L. Kelly and C. C. Lochbaum, "Speech synthesis," in *Proc. Fourth Int. Congr. Acoustics*, paper G42, pp. 1–4, Copenhagen, Denmark, Sept. 1962.

[12] 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.