Generalized Source-Filter Structures for Speech Synthesis

Matti Karjalainen and Tuomas Paatero

Helsinki University of Technology
Laboratory of Acoustics and Audio Signal Processing
P.O. Box 3000, FIN-02015 HUT, Espoo, Finland
matti.karjalainen@hut.fi

Abstract

In this paper we discuss various digital filter principles as models for synthetic speech generation. Warped linear prediction (WLP) and frequency-warped filters have been introduced earlier as a method to reduce the filter order in high-quality wideband speech synthesis. In addition to analyzing WLP and frequency-warped filters we introduce new related structures and techniques for arbitrary frequency resolution allocation. Kautz filters can be considered as generalized structures for pole-zero modeling. This study focuses on residual-excited synthesis and diphone-oriented reconstruction of speech signals. Control strategies for text-to-speech synthesis are discussed briefly.

1. Introduction and Motivation

Generation of speech signals in speech synthesis is typically based on a source-filter model, waveform concatenation of prerecorded signal samples, or a combination of these methods. Except in time domain overlap-add techniques such as PSOLA [1] or microphonemic synthesis [2], a (digital) filter model and an appropriate excitation signal is needed. Among desired features for such synthesis are:

1. low-order filter structure that is computationally efficient[1],
2. a natural set of filter control parameters to realize dynamic transitions within and between phonemes,
3. a systematic method to derive the filter control parameters from recorded speech, and
4. a systematic method to analyze a compact excitation signal or model, based on recorded speech.

In early synthesis methods the source-filter models of speech production were more or less hand-tuned both for the controlling parameters and excitation signal model. It was found that several filter approaches are theoretically equivalent but may have pros and cons from implementation or synthesis control points of view. Cascaded second-order blocks and parallel structures are traditional examples thereof. Combinations and hybrids of these have been successful also [3, 4].

In early speech synthesis the control parameters were, if possible, formant frequencies and other formant-related parameters, more or less fulfilling the second requirement above. The last two conditions were only met when linear prediction [5] was available as a source-filter modeling technique. Parametric glottal models [6] may yield a good overall quality, but individual features of a particular speaker are difficult to model in detail except using an inverse-filtered excitation analyzed from real speech samples.

In this paper we are interested in source-filter synthesis structures that meet the four requirements stated above. We start from a viewpoint of linear prediction as a speech modeling technique, presenting a short discussion of traditional filter structures. A generalization to frequency-warped linear prediction (WLP), as specified in [7] and [8], is discussed from excitation, implementation, and control viewpoints. As a further generalization, Kautz filters are presented in the context of speech synthesis, as a means to design pole-zero filters with arbitrary focusing of frequency resolution. The problems of modeling the excitation signals as well as parametric control strategies for source-filter speech synthesis are discussed briefly.

[1] Only digital filter implementations are considered here.

2. Linear Prediction and Traditional Filter Structures

Linear prediction and related source-filter modeling of speech signals is one of the most important techniques in speech processing [5]. A general discrete-time linear and time-invariant (LTI) source-filter signal model is $Y(z) = S(z)H(z)$, where $S(z)$ is the source signal, $H(z)$ is the filter, and $Y(z)$ is the resulting signal. A general pole-zero filter $H(z)$ has the form

$$H(z) = \frac{\sum_{i=0}^{M} b_i z^{-i}}{1 - \sum_{i=1}^{N} a_i z^{-i}}. \quad (1)$$

Linear prediction is an efficient technique to find optimal parameter values a_i for an all-pole version of (1), i.e., $H(z) = G/(1 - \sum_{i=1}^{N} a_i z^{-i})$, where the numerator has only a gain factor G. A standard technique of obtaining optimal filter coefficients $\tilde{a}_i$ is to compute autocorrelation coefficients of the speech signal under study and to solve the normal equations constructed from these coefficients.

In practice the application of linear prediction in text-to-speech synthesis contains the following subproblems. The selection of the order N of an all-pole model works to allocate two poles per 1 kHz of bandwidth plus two, being able to model the formants and the general spectral shape. For high-quality wideband speech for sample rates of 22–48 kHz the number of a_i parameters becomes inconveniently high. This will be discussed below in the context of warped linear prediction.

The next problem is the control of filter parameters in order to realize appropriate transitions within and between phonemes. If the desired parameters are know at specific time moments, the problem becomes how to interpolate them properly between these values. Techniques that are known to be useful are using reflection coefficients related to a lattice filter formulation of the all-pole filter, log-area ratio (LAR) parameters derived from them, or line spectrum frequencies (LSF). These guarantee that the filter remains stable while interpolating between two stable filters. Differences between these methods exist but in practice they are not very prominent.

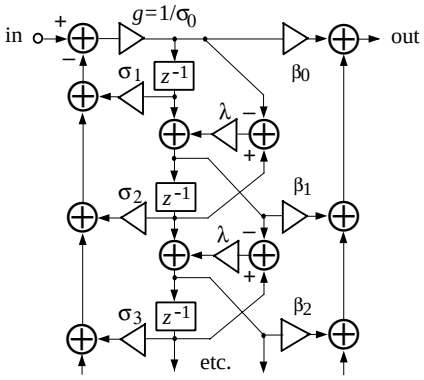

Figure 1: Realizable synthesis filter structure for WLP.

The third subproblem with linear prediction, as with any residual excited source-filter model, is the realization of filter excitation. Since the excitation is essentially a non-minimum-phase signal, a low-order yet precise parametric model is difficult to find. A straightforward technique is to make an inventory of inverse-filtered residuals, sampled from a representative set of phoneme contexts. Compact coding of this set may be applied, such as vector quantization, but for highest quality such a codebook of excitations still takes quite an amount of memory. This problem will be discussed later in Section 5.

3. Warped Linear Prediction

The first systematic formulation of warped linear prediction for speech signals was presented by Strube [9]. Later, Laine *et al.* [7] have studied various formulations of efficient WLP. The idea of a warped frequency scale and related resolution is based on using allpass sections instead of unit delays in DSP structures [2], i.e.,

$$\tilde{z}^{-1} = D_1(z) = \frac{z^{-1} - \lambda}{1 - \lambda z^{-1}} \qquad (2)$$

where λ, $-1 < \lambda < 1$, is a warping parameter and $D_1(z)$ is a warped (dispersive) delay element. With a proper value of λ, the warped frequency scale shows a good match to the psychoacoustically defined Bark scale [10], thus optimizing the frequency resolution from the point of view of auditory perception. For example, with a sampling rate of 22 kHz, Bark warping is obtained using λ = 0.63. WLP analysis is easily realized by modifying only the autocorrelation computation using a version where unit delays are replaced by allpass sections. The same holds for inverse filtering to obtain the residual (excitation) signal. The synthesis filter, however, cannot be realized in such a simple manner since in recursive structures the replacement of Eq. (1) results in delay-free loops. Techniques to avoid this problem are discussed, e.g., in [11]. The filter structure shown in Fig. 1 has been used in our WLP synthesis experiments. The original (warped) denominator coefficients are mapped to new coefficients σ_i that are used as feedback coefficients. Otherwise, the WLP analysis and synthesis techniques are the same as with ordinary linear prediction.

The advantage gained when using Bark warping is that in wideband synthesis the filter order can be reduced remarkably without sacrificing the frequency resolution at low frequencies.

[2] A systematic orthonormal formulation of frequency warping can be given by Laguerre functions.

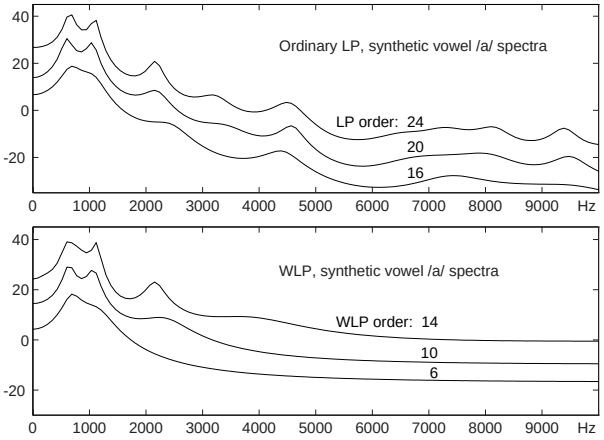

Figure 2: LP and WLP spectra of vowel /a/ for different filter orders.

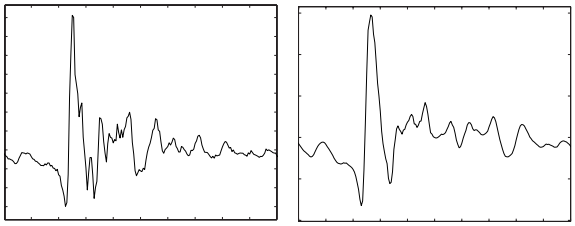

Figure 3: Residual signals for the LP (left) and WLP (right) cases of Fig. 2.

At high frequencies the spectral resolution is worse, nevertheless this is exactly how the human hearing functions.

It turns out that WLP helps reducing the filter order substantially if the sampling rate is high, such as 22 or 44 kHz [8]. For example for 22 kHz the ordinary LP yields good quality with orders of 20–24 while WLP works comparably with orders of 10–14. Approximately same WLP orders are valid also for a 44 kHz sample rate. Figure 2 shows synthesis filter responses for a vowel spectrum (Finnish /a/) using ordinary LP and WLP.

Experimentally we have found that the success of low-order WLP is based on representing the necessary critical band frequency resolution by WLP synthesis filter and the necessary temporal fine-structure for individual voice by the inverse-filtered residual. Figure 3 depicts the residual signal for the case of Fig. 2 for WLP and ordinary LP. Both the excitation and the filter parameters can be interpolated or approximated successfully so that the reconstructed signal waveform is competitive in quality with high-order ordinary LP. This emphasizes the importance of the temporal fine-structure of signals in speech synthesis.

4. Kautz Filters

The lowest order rational functions, square-integrable and orthonormal on the unit circle, analytic for $|z| > 1$, are of the form [12]

$$G_i(z) = \frac{\sqrt{1 - z_i z_i^*}}{z^{-1} - z_i^*} \prod_{j=0}^{i} \frac{z^{-1} - z_j^*}{1 - z_j z^{-1}}, \quad i = 0, 1, \ldots, \qquad (3)$$

defined by any set of points $\{z_i\}_{i=0}^{\infty}$ in the unit disk. Functions (3) form an orthonormal set which is complete, or a base, with a moderate restriction on the poles $\{z_i\}$ [12]. The corresponding time functions $\{g_i(n)\}_{i=0}^{\infty}$ are impulse responses or inverse

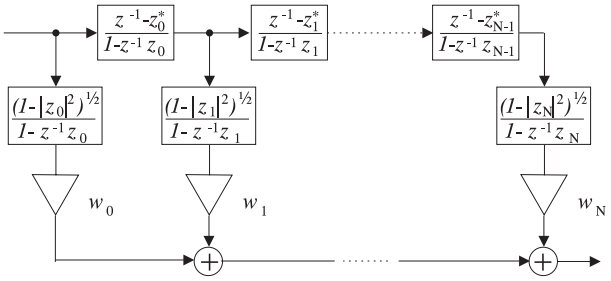

Figure 4: The Kautz filter. For $z_i = 0$ in (3) it degenerates to an FIR filter and for $z_i = a, -1 < a < 1$, it is a Laguerre filter where the tap filters are replaced by a common pre-filter.

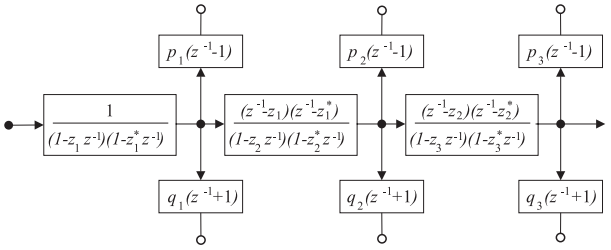

Figure 5: One realization for producing real Kautz functions from a sequence of complex conjugate pole pairs.

z-transforms of (3). This implies that a basis representation of any causal and stable discrete-time signal or LTI system is obtained as its Fourier series expansion with respect to the time or frequency domain basis functions. These generalizations of z-transform and convolution sum representations provide linear-in-parameter models for signals and systems.

A Kautz filter [13] is a finite weighted sum of functions (3), which clearly reduces to a transversal structure of Fig. 4. The filter structure is completely determined by a pole set $\{z_i\}_{i=0}^{N}$ and a weight vector $\mathbf{w} = [w_0 \cdots w_N]^T$. We define the filter or model order to be $N + 1$.

A Kautz filter produces real tap output signals only in the case of real poles. However, from a sequence of real or complex conjugate poles it is always possible to form real orthonormal structures. From the infinite variety of possible solutions it is sufficient to use the intuitively simple structure of Fig. 5, proposed by Broome [14]: the second-order section outputs of Fig. 5 are *orthogonal* from which an orthogonal tap output pair if formed. Normalization terms are completely determined by the corresponding pole pair $\{z_i, z_i^*\}$ and are given by

$$\begin{aligned} p_i &= \sqrt{(1-\rho_i)(1+\rho_i-\gamma_i)/2} \\ q_i &= \sqrt{(1-\rho_i)(1+\rho_i+\gamma_i)/2} \end{aligned} \tag{4}$$

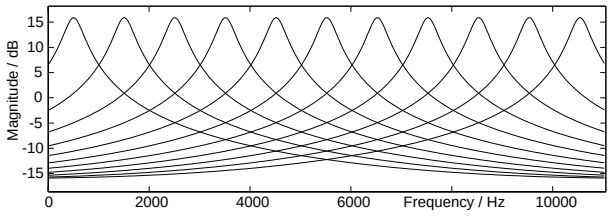

Figure 6: Tap output magnitude responses of a Kautz structure with complex poles to demonstrate the parallel formant synthesizer behavior of the filter. Sample rate is 22 kHz.

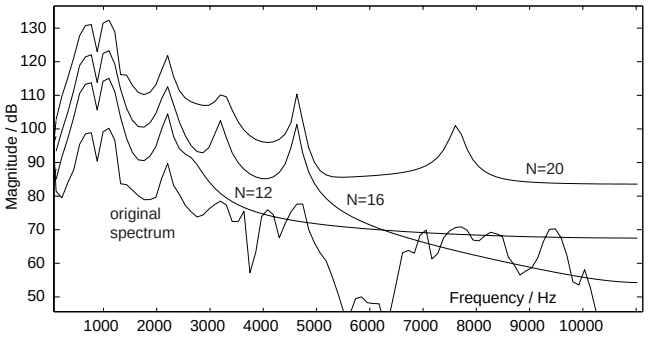

Figure 7: Spectral modeling of vowel /a/ with Kautz filters: original spectrum and magnitude responses of orders 12, 16, and 20.

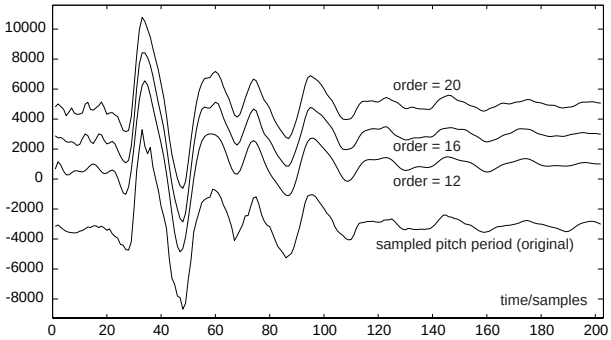

Figure 8: Temporal modeling of vowel /a/ with Kautz filters: one pitch period and filter impulse responses of orders 12, 16, and 20.

where $\gamma_i = -2RE\{z_i\}$ and $\rho_i = |z_i|^2$ can be recognized as corresponding second-order polynomial coefficients. The construction works also for real poles but we use an obvious mixture of first- and second-order sections, if needed.

Kautz filter design can be seen as a two-step procedure involving the choosing of a particular Kautz filter (i.e., the pole set) and the evaluation of the corresponding filter weights. For the latter, and a given target response $h(n)$ or $H(z)$, we use simply the Fourier coefficients, $c_i = (h, g_i) = (H, G_i)$, which can be evaluated by feeding the signal $h(-n)$ to the Kautz filter and reading the tap outputs $x_i(n) = G_i[h(-n)]$ at $n = 0$: $c_i = x_i(0)$. This implements convolutions by filtering and it can be seen as a generalization of rectangular window FIR design. Especially in low-order modeling, however, the most essential part in Kautz filter design is the choosing of poles. There are many methods that can be used in search for suitable poles, including all-pole or pole-zero modeling, sophisticated guesses, and random or iterative search, but here we name just two, in a sense opposite strategies. A Kautz filter impulse response is a weighted superposition of damped sinusoidals, which provide for direct tuning of a set of resonant frequencies and corresponding time constants. As a contrast to manual tuning, we have adopted a method proposed originally to pure FIR-to-IIR filter conversion [15], to the context of Kautz filter pole optimization. It resembles the iterative Steiglitz-McBride method of pole-zero modeling, but it genuinely optimizes (in the LS sense) the pole positions of a real Kautz filter, producing unconditionally stable and (theoretically globally) optimal pole sets for a desired filter order.

A basic property of Kautz filters is shown in Fig. 6 where the tap output responses of the structure of Fig. 4 with complex-valued poles, having the same radius and evenly distributed pole angles, are plotted. It shows that the structure forms a parallel formant filterbank. Although a real-valued filter of Fig. 5 does exhibit slightly different behavior, it is evident that the filter is inherently well suited to model the formant behavior of speech signals. The power of this formalism is that, if the poles are properly positioned, a least squares optimal fit of the tap coefficients to a response to be modeled is as easy as with FIR filters.

Figure 7 depicts the spectral modeling ability of vowel /a/ for Kautz filter orders 12, 16, and 20. For low orders only the high spectral peaks are modeled accurately, but for higher orders the match is better, especially when listening to the synthesized signals. Figure 8 illustrates the same cases from a temporal modeling viewpoint. It is interesting to notice that the 'non-minimum phase' behavior of the pitch period is modeled properly with Kautz filters, and the resulting synthesized sound includes individual sound qualities of the person who uttered the original sample.

To use the designed Kautz filter for synthesis, simply an impulse train of the desired fundamental frequency id fed in, and the filter generates a sequence of pitch periods, such as in Fig. 8. Thus the control of pitch is very easy.

5. Excitation and Control Strategies

The excitation of a source-filter synthesis model can be formed in several ways. For highest-quality, individual voice synthesis is in general possible only by applying inverse filtered residual or its approximation, taken from a similar spoken utterance.

Selection of a maximally similar spoken unit from an inventory of speech samples and using such excitation waveform together with corresponding filter coefficients yields a perfect reconstruction for the time positions where the sampling was applied. Careful interpolation of both excitation waveform and the filter coefficients between such temporal positions can work well, as was demonstrated for WLP in [8].

Sampled pitch periods of residual signals take relatively much memory, unless properly coded or compressed. In many cases this is not a problem, however, when the synthesis software runs on a modern desktop computer. In portable devices, or when the model data must be transmitted in a narrowband transmission channel, this is more critical.

The following principles are useful to compress the excitation data: (a) Codebook designs; any technique to form a compact quantized codebook; (b) Multipulse excitation design, i.e., searching for an optimal nonuniform sampling to represent the excitation; (c) From an auditory point of view, a possible idea is to lowpass filter the excitation at about 1.5 kHz and decimate the low-frequency residual. The 1.5 kHz high-passed residual can be rectified and the envelope lowpassed and decimated. The problem of coding the envelope resembles the problem in multipulse excitation; to find a representation which after the reconstruction yield best synthesis quality.

For controlling the synthesis models in text-to-speech synthesis in the context of WLP synthesis, neural networks [8] and unit parameter selection from a codebook of database items have been experimented. Neural networks were found useful only in low-to-medium quality synthesis since the parametric accuracy was not particularly good. The latter method, with a codebook of excitations and filter parameters, was found more appropriate. Control strategies for Kautz filter synthesis have not yet been studied.

6. Summary and Conclusions

This paper has discussed the possibilities to generalize source-filter models for speech synthesis. Frequency-warped filters and linear prediction are found as a technique to reduce the filter order for high sampling rates substantially by utilizing the auditory frequency resolution. Kautz filters are introduced as a further generalization of rational functions for designing synthesis filters that are able to model the excitation properties as well.

7. Acknowledgments

This study is related to Tekes project "Development of Audio-visual Speech Synthesis". The work of M. Karjalainen and T. Paatero has been supported primarily by the Academy of Finland, related to project "Sound Source Modeling".

8. References

[1] Moulines E., and Carpenter F., "Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones," *Speech Communication*, 9(5/6):453–467, Dec. 1990.

[2] Lukaszewicz K., Karjalainen M., "Microphonemic Method of Speech Synthesis," *Proc. of IEEE ICASSP-87*, Dallas, 1987.

[3] Klatt D., "Software for Cascade/Parallel Formant Synthesizer," *J. Acoust. Soc. Am.* vol. 67:971-995.

[4] Laine U. K., "PARCAS, a New Terminal Analog Model for Speech Synthesis," *Proc. of IEEE ICASSP-82*, Paris 1982.

[5] Markel J. D., and Gray A. H., *Linear Prediction of Speech*, Springer Verlag, New York, 1976.

[6] Fant G., Liljencrants J., and Lin Q. C., "A Four-Parameter Model of Glottal Flow," *Speech Transmission Lab. Quarterly Progress and Status Report*, KTH, Stockholm, 1985, no. 4, pp. 1–13.

[7] Laine U. K., Karjalainen M., and Altosaar T., "Warped Linear Prediction (WLP) in Speech and Audio Processing," *Proc. IEEE ICASSP -94*, Adelaide, Australia, 1994.

[8] Karjalainen M., Altosaar T., and Vainio M., "Speech Synthesis Using Warped Linear Prediction and Neural Networks," *Proc. IEEE ICASSP-98*, 1998.

[9] Strube H. W., "Linear Prediction on a Warped Frequency Scale," *J. Acoust. Soc. Am.*, vol. 68, no. 4 (1980).

[10] Smith, J. O., and Abel, J. S. "The Bark Bilinear Transform," *Proc. IEEE ASSP Workshop*, New Paltz, 1995.

[11] Karjalainen M., Härmä A., and Laine U.K., "Realizable Warped IIR Filters and Their Properties", *Proc. IEEE ICASSP-96*, Munich, 1996.

[12] J. L. Walsh, *Interpolation and Approximation by Rational Functions in the Complex Domain*, 2nd Edition. American Mathematical Society, Providence, Rhode Island, 1969.

[13] W. H. Kautz, "Transient Synthesis in the Time Domain", *IRE Trans. Circuit Theory*, vol. CT–1, pp. 29–39, 1954.

[14] P. W. Broome, "Discrete Orthonormal Sequences", *Journal of the Association for Computing Machinery*, vol. 12, no. 2, pp. 151–168, 1965.

[15] H. Brandenstein and R. Unbehauen, "Least-Squares Approximation of FIR by IIR Digital Filters", *IEEE Trans. Signal Processing*, vol. 46, no. 1, pp. 21–30, 1998.