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 wide-band 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 pre-recorded 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 efficient1,
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.

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

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

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 \( a_i \) 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 wide-band speech for sample rates of 22-48 kHz the number of all-pole 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 known 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.
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, i.e.,

\[ \frac{z^{-1}}{\lambda} = D_1(z) = \frac{z^{-1} - \lambda}{1 - \lambda z^{-1}} \]  

where \( \lambda, -1 < \lambda < 1 \), is a warping parameter and \( D_1(z) \) is a warped (dispersive) delay element. With a proper value of \( \lambda \), 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 \( \lambda = 0.63 \). 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 \( \sigma_i \), 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.

4. Kautz Filters

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

\[ G_i(z) = \frac{1 - z^{-1} \zeta_i^*}{z^{-1} - \zeta_i^*} \prod_{j=0}^{i-1} \frac{z^{-1} - \zeta_i^*}{1 - \zeta_i z^{-1}}, \quad i = 0, 1, \ldots \]  

defined by any set of points \( \{\zeta_i\}_{i=0}^{\infty} \) in the unit disk. Functions (3) form an orthonormal set which is complete, or a base, with a moderate restriction on the poles \( \{\zeta_i\} \) [12]. The corresponding time functions \( \{g_i(n)\}_{i=0}^{\infty} \) are impulse responses or inverse
where the tap filters are replaced by a common pre-filter. From the in-finite conjugate poles it is always possible to form real orthonormal filters, 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

\[
\begin{align*}
\rho_i &= \sqrt{(1 - \rho_i)(1 + \rho_i - \gamma_i)} / 2, \\
\gamma_i &= -2RE\{z_i\} \text{ and } \rho_i = |z_i|^2 \text{ can be recognized as corresponding second-order polynomial coefficients. The construction works also for real poles but we use an obvious mixture of first- and second-order sections, if needed.}
\end{align*}
\]

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 least squares sense) the pole positions of a real Kautz filter, producing unconventionally 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 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 Audiovisual 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