EXPLOITING IMPROVED PARAMETER SMOOTHING WITHIN A HYBRID CONCATENATIVE/LPC SPEECH SYNTHESIZER

Erhard Rank*

Institute of Communications and Radio-Frequency Engineering
Vienna University of Technology, Gusshausstrasse 25/E389, A–1040 Vienna, Austria
Email: erank@nt.tuwien.ac.at

Abstract

We depict the interpolation strategies for the concatenation of inventory demisyllables in our hybrid concatenative/LPC speech synthesizer. Inventory elements for vowels and nasals are cut in the steady state of the phoneme. Concatenating elements in the synthesis stage requires smoothing of spectral content and energy to avoid annoying discontinuities in these parameters, which is of vital importance for the quality of synthesized speech. The hybrid synthesizer concept allows the application of smoothing algorithms uncommon to pure time domain synthesis. Using the smoothing algorithm also the occurrence of musical tones due to the LPC filter resonances at the beginning of pauses can be suppressed.

1 INTRODUCTION

The good quality of state-of-the-art speech synthesizers is mainly due to the use of concatenative synthesis: synthesis by concatenation of prerecorded speech segments generally yields more natural speech than, e.g., synthesis with Klatt-type models. However, concatenative synthesis leaves us with the task of parameter smoothing around many concatenation points to avoid discontinuities of loudness and spectral content, in particular for general purpose synthesizers that have to use short inventory elements to facilitate synthesis of any desired utterance.

Two of the commonly used inventory concepts—namely the use of diphone and demisyllable elements—involves concatenation inside the vowel regions where sudden changes of loudness or spectral content are clearly perceivable and very annoying to the listener. Strategies to cope with the problem are, e.g., the extension of the inventory with context-dependent elements and use of the best matching elements at a time [1] or the use of larger inventory units [2], which makes discontinuities less frequent but increases inventory size considerably. On the other hand, with appropriate smoothing it is possible to avoid audible steps at the concatenation points while keeping the inventory size modest.

Due to the non-parametrized representation of inventory speech segments in many concatenative synthesizers parameter smoothing becomes difficult, especially when sophisticated, articulatory-related smoothing strategies are to be applied. The derivation of parameters from the inventory representation which lend themselves for smoothing, like formant trajectories, and the inverse operation—modifying the inventory elements according to the interpolated parameter values—is far from trivial.

In this paper, we describe parameter smoothing for a synthesis method that combines the benefits of a concatenative synthesis concept with those of source-filter modelling. First, the design of the hybrid synthesis method is presented. Then the properties of the signals and system elements concerned with parameter smoothing are specified and, in the following, a simple and efficient smoothing algorithm is described in detail and examples for the efficacy of this algorithm are presented.

2 THE HYBRID SYNTHESIS CONCEPT

Several recent approaches to high-quality template-based speech synthesis apply a hybrid concatenative/LPC (linear predictive coding) model instead of pure time-domain concatenation for the synthesizing stage [3, 4]. The advantages of residual excited linear prediction (RELP) synthesis—it allows for better modifications of prosodic parameters while keeping the high quality of concatenative synthesis—are well known [5, 6]. The hybrid approach is not only well suited for manipulating f0 and duration [7], but also provides the means for a simple and efficient method for parameter smoothing at concatenation points resembling the physics of the speech production process.

The synthesis concept described in the following is implemented in a fully functional software synthesizer used for phoneme-to-speech conversion in the Vienna concept-to-speech system (VIECTOS, [8]).

The synthesizer is based on demisyllable elements recorded from a male speaker. The sampling frequency of the inventory speech is 16 kHz. For creating an inventory, pitch periods are labelled semi-automatically in the recorded segments and a pitch-synchronous LPC analysis...
is performed [9]. Both the coefficients for the LPC filter and the corresponding residual signal are stored in the inventory.

During the LPC analysis, the peak of the residual of a voiced frame is centered within the frame. Thus, for each voiced frame, the residual vector $x_{\text{res}}$ contains a main pulse in the center and the energy typically decays towards the beginning and the end whereas, for unvoiced frames, the residual is more noise-like. In terms of temporal energy distribution each voiced residual frame thus resembles a single glottis pulse. A typical example for a voiced residual frame and the corresponding LPC filter transfer function is depicted in fig. 1.

The LPC analysis/synthesis filter is realized as a lattice filter. The lattice filter is a discrete-time model for wave propagation in a one-dimensional waveguide with varying wave impedance. The length of the LPC filter is chosen $N = 18$, which relates to the length of the human vocal tract equidistantly sampled with a sampling distance corresponding to the 16 kHz sampling frequency [9]. The lattice filter implementation of the LPC filter thus provides a physical model of the vocal tract and corresponds to a acoustic tube model in the following way: the reflection coefficients $k_m$ of the lattice structure can be directly related to log-area-ratios (LAR) for a tube model by

$$\text{LAR}_m = \log_{10} \left( \frac{1 + k_m}{1 - k_m} \right), m = 1 \ldots N.$$  

LARs are the logarithmic ratios of cross-section areas for the sections of a tube model.

The so defined synthesis model hence involves an LPC filter that directly models vocal tract properties and a residual signal resembling glottis pulses to some extent. All prosodic manipulations (fundamental frequency, segmental duration) can be done in the residual domain [7], and, beyond that, the model allows manipulations of some voice quality parameters at the residual signal and the LPC filter [10].

### Figure 1: Typical residual and LPC filter transfer function for a frame in a voiced phoneme. This example is from the phoneme /a:/.

The energy of the residual pulse is concentrated in the frame center.

### Figure 2: Transition between the vowels /a:/ and /e:/ generated by interpolation of LAR coefficients. On the left side, the tube model calculated from the reflection coefficients of the LPC filter is shown for both vowels. The spectra on the right side show the transfer functions of these LPC filters (solid lines) and from three intermediate filters (dotted lines). Note: the demisyllable synthesizer does not perform transitions between different phonemes in this way, interpolation is applied at junctions of nominally the same phoneme only.

### 3 SMOOTHING ALGORITHM

The synthesis model splits the speech signal into LPC coefficients and residual signal, so distinct smoothing strategies can be applied for both. Since the LPC analysis is pitch synchronous, waveform interpolation could be used for the residual, but we found that the residual holds much of the frame-to-frame variability of natural speech which would be lost if a series of residual frames was strictly interpolated. So for the residual signal only the amplitude is scaled in order to smooth the energy $E_{\text{out}}$ of the output signal.

For interpolation of LPC filter coefficients a vast number of possibilities exist (see, e.g., [11]). Some of these have the drawback of possible instability of interpolated filters, whereas others are stable with different performance depending on the distortion measure applied. To achieve minimal spectral distortion interpolation of line spectral frequencies (LSFs) can be applied [12], which results in a smooth transition of the formants. Anyhow, since the LAR representation of the LPC filter coefficients describes a tube model of the vocal tract, interpolating LARs corresponds to a smooth transition from starting position to end position for each part of the vocal tract (lips, tongue, . . .). Thus a seamless transition of the LPC spectra at the junction of two parts of a voiced phoneme originating from different inventory elements can be achieved by interpolating LARs. The interpolation of LARs always yields stable LPC filters.

An example of LAR interpolation is shown in fig. 2. On the left side the original state and the final state of
Figure 3: Example of synthesizer output without and with smoothing applied. The figure shows the time signal and the wideband spectrogram for the concatenation of the two demisyllables <tso-> and <-ox> without (top) and with (bottom) interpolation. The concatenation point is marked by the vertical line in the signal plot at about 150ms. For the interpolated version the region of interpolation is shaded in the signal plot. The step of the second formant in the non-interpolated case is gently smeared in the interpolated case.

the corresponding tube model is shown and on the right-hand side the transition between the LPC spectra is plotted. It has to be noted that a transition like this from one phoneme to another would not usually be performed in the synthesizer. Only for concatenations of a demisyllable ending and another demisyllable starting with nominally the same phoneme and transitions of voiced phonemes towards pauses (or vice versa), interpolation of the LPC filter can be applied with satisfactory results. For transitions between different vowels, dedicated segments are provided in the inventory.

As noted before, the residual in the interpolated regions is scaled according to the gain of the interpolated LPC filter, such that the energy of the resulting output signal is smoothed:

$$x_{res} = \left( \frac{E_{out} \cdot \sum_{m=1}^{N} (1 - k_m^2)}{E(x_{res})} \right)^{1/2} x_{res}$$

where $x_{res}$ is the residual vector before resp. after scaling, $E(x_{res})$ the energy of $x_{res}$, $k_m$ the modified reflection coefficients for the LPC filter, and $E_{out}$ the desired energy of the output signal. In this way a smooth energy transition and often—but not always—also a smooth envelope of the speech signal is obtained. With this scaling amplitude mismatches in the inventory are corrected also.

Fig. 3 shows the effect of the smoothing algorithm on the waveform and the spectrogram of the synthesized
speech. The discontinuities inside the vowel in the frequency range above 1kHz, especially the mismatch at the second formant visible in the non-interpolated case, are eliminated by the interpolation.

Sometimes it is also desirable to smooth transitions of voiced phonemes towards pauses\(^1\). If the LPC filter is kept running with the recent coefficients and little or no excitation signal this often results in the emergence of musical tones at the resonances of the LPC filter. So besides the attenuation of the residual signal a damping of the LPC filter resonances is needed. For the lattice structure filter this is achieved by a multiplication of the coefficients \(k_m\) with a factor \(f < 1\). The corresponding process in the LAR domain is a transition towards a uniform tube model, i.e., towards the neutral state of the vocal tract.

Due to the specific pitch-synchronous LPC analysis, both the LPC filter interpolation and the residual signal scaling can be performed on a frame by frame basis. Because the energy of the residual is concentrated in the frame center, the effect of parameter switching at the frame borders is negligible. The mismatches of the excitation signal\(^2\) are furthermore smoothed by the memory of the LPC filter.

4 CONCLUSION

For concatenative speech synthesis appropriate smoothing at concatenation points is an alternative to the use of large size inventories. Nevertheless, accessing the optimal smoothing parameters remains a problem, especially for synthesis methods using a pure time-domain signal representation in the inventory.

We described the smoothing strategies for a hybrid concatenative/LPC model synthesizer that uses residual-excited linear predictive (RELP) synthesis. The distinction between voice source (residual) and vocal tract (LPC filter) as separate parts of the speech production process is a promising concept for future synthesis systems.

Due to the realization of the LPC filter as a lattice filter, simple yet effective strategies for spectral smoothing can be implemented while excitation with the original residual pulses keeps the high quality associated with concatenative synthesis. The application of this smoothing strategy for the concatenation of all vowels and nasals in our demisyllable based synthesizer resolves the problem of annoying discontinuities of loudness and spectral content encountered for the concatenation of allophonic variants of a phoneme in the inventory.

Utilizing the methods for smoothing also the occurrence of musical tones due to the LPC filter resonances at the begin of pauses can be suppressed.

\(^1\)In our synthesizer phrasing pauses as well as plosive pauses are generated by zeroing the excitation of the LPC filter.

\(^2\)Further mismatches of the excitation signal occur due to the pitch modification algorithm.

REFERENCES


