A Wavelet-Domain PSOLA Approach

M. Holzapfel* **, R. Hoffmann** and H. Höge*

* Siemens Corp. Otto Hahn Ring 6, D-81739 Munich, Germany
** Institute for Technical Acoustics, Technical University of Dresden, D-10162 Dresden, Germany
Martin.Holzapfel@mchp.siemens.de

Abstract
A basic problem in concatenative speech synthesis are discontinuities at the concatenation points. The units which are produced by different (independent) articulatory movements differ in their spectral characteristics even if their phonetic context is carefully chosen. This paper describes a wavelet transform of the spectrum of the speech concatenated within the PSOLA algorithm. This multiresolution analysis separates the following perceptive important spectral characteristic: the intrinsic pitch resulting in a fine-ripple of the spectrum, articulatory movements typically resulting in formant-structures and the global spectral tilt.

In the wavelet domain each of this characteristic can be analysed and manipulated separately in a consistent and completely non parametric way. Optimised concatenation points can easily be located. Remaining spectral irregularities can be adjusted efficiently, resulting in clear and naturally sounding synthetic speech.

Dyadic filter banks are a computational efficient implementation of the presented transform.

1. Introduction.

To achieve high quality in concatenative speech synthesis the representation of the speech signal has to fulfil two requirements:
- preserving all information and naturality contained in the prerecorded speech samples,
- allowing easy and consistent manipulation of the spectral characteristics to avoid audible distortion at the concatenation points.

For little prosodic manipulation TD-PSOLA [1] is an excellent way to preserve naturality of a prerecorded voice in concatenative speech synthesis but does not enable spectral manipulation. FD-PSOLA [2] conserves as well all information contained in speech originally but further allows to correct the remaining intrinsic pitch. Smoothing spectral discontinuities in the FD-representation leads to poor results, as the scope is limited to single frequency bands and fails to adjust complex shapes.

Parametric approaches like the "source filter"-LPC-PSOLA [3] or sinusodial models [5] allow direct control of the spectral shape of the speech signal but tend to loose detail information and so decrease naturality.

Wavelet-Domain PSOLA avoids both shortcomings and adds a multiresolution point of view. The spectrum of the speech in each Hamming window ("PSOLA-bell") is decomposed to the translates and dilates of a complex pattern, the wavelet function.

In section 2 we will show that by choosing a suitable wavelet function, perceptually relevant parts of the spectrum can be separated in different levels of resolution and the corresponding wavelet coefficients. The algorithms presented in section 3 uses this decomposition to find optimised concatenation points and to adjust spectral shapes in concatenation. Section 4 describes the embedding of those algorithms in our TTS system. Section 5 proposes an dyadic filter bank implementation of the ideal transform developed in chapter 2.

2. Design of a wavelet transform, suitable for spectral representation.

We focus on the spectrum of the pitch-synchronous windows of double pitch-length used in the wellknown PSOLA algorithm[4]. Signal processing proceeds in several stages: first the Hamming windowing, second a FFT, third a computation of spectral magnitude and phase, finally an analysis of the spectrum is desirable.

The spectrum \(s(f)\) is decomposed to its inner products \(c_{ia}\) by the translates (denoted by the index \(i\)) and dilates (index \(a\)) of a wavelet function \(\Psi^i_a(f)\) and a lowpass residuum \(\xi(f)\).

\[
s(f) = \sum_i \sum_a c_{ia} \cdot 2^{-a} \Psi^i_a \left( \frac{f - i \cdot 2^a}{2^a} \right) + \xi(f) \quad (1)
\]

\[
c_{ia} = \int_{-\infty}^{\infty} s(f) \cdot 2^{-a} \Psi^i_a \left( \frac{f - i \cdot 2^a}{2^a} \right) df \quad (2)
\]
The spectral patterns of major interest are: periodicities, local energy peaks and global tilt. A standard wavelet popular in neurology and image analysis [7] [8] is modified to match those patterns. There are hints that human perception is organised in a similar way [6]. The real part of this wavelet is given by (3). For the complex extension see (4) and (5).

$$\psi(f) = e^{\left(1 - \left(\frac{f}{\sigma}\right)^2\right)} \cdot e^{-\frac{1}{2} \left(\frac{f}{\sigma}\right)^2}$$  \hspace{1cm} (3)

**Figure 1: Graph of the mexican hat function**

The mexican hat function is well tuned to the three characteristics of interest:

- Choosing a $\sigma$ of $\sqrt{3}$ the position of the sideband minima are at +/-1, the shape of the “mexhat” is very similar to a windowed cosine-function oscillating at the Nyquist-frequency. Those fine ripples are mainly caused by the pitch intrinsic to the “PSOLA-bell” [2]. The wavelet coefficients $c_{ia}$ at the zeroth and first dilation level are dominated by the influence of the voiced excitation by the vocal cords. Talking in terms of a source filter model of speech production, one would talk of those wavelet coefficients $c_{ia}$ representing the source component of speech (see top line of Figure 3 for example).

- In the spectral envelope, the inner product by this function $c_{ia}$ indicates local energy maxima of distinctive mid frequencies and bandwidths. These maxima in the spectrum are caused by vocal tract resonance frequencies due to articulator’s positions. Even in the absence of formants, the perceptually important shape of the spectrum is explicitly controlled by the wavelet parameters $c_{ia}$ corresponding to the middle-range dilates of the wavelet function (see mid line of Figure 3).

- The low levels of dilation matches the overall spectral tilt (indicated by its similarity to the centre part of the extremely dilated wavelet function). (see the base line in Figure 3)

This wavelet function is well localised in time and frequency. For the above mentioned choice of $\sigma$ the function spaces spanned by the translates and dilates of the wavelet (1) (2) are very close to orthogonal.

We are not interested in indicating strictly phaselocked cosine structures, but in arbitrary periodicities, maxima and bandwidths. So the analysis pattern is extended by its phase-shifted version as complex part of the wavelet function (corresponding to sine and cosine as real and imaginary part of the Fourier transform). This phase-shifted version can directly be computed by the Hilbert transform of the real wavelet function $\psi(f)$ defined in (3) [10]. The complex wavelet function $\Psi(f)$ is defined as:

$$\Psi(f) = \psi(f) + j\mathcal{H}\{\psi(f)\}$$  \hspace{1cm} (4)

**Figure 2: Formula and graph of the real and imaginary part of the wavelet function (the mex-hat (doted line), see figure 1, and its Hilbert transform (solid line))**

The constant $c$ in formula 3 is chosen to normalise the complex wavelet function:

$$\int_{-\infty}^{\infty} \Psi(f)\Psi^*(f) df = 1$$  \hspace{1cm} (5)

**Figure 3: 256 point spectrum of a vowel in multiresolution performed by the filter structure presented in section 5. The right column shows the spectrum reconstructed from the wavelet coefficients $c_{ia}$ of the dilation levels a: 1 (top line), [1; 3] mid line and [1; 5] base line. The left column shows the corresponding lowpass residuum $\xi(f)$. The wavelet coefficients contain the difference signal of two adjacent lowpass residuum signals. In this example the wavelet coefficients of level 4 and 5 capture the formant structure.**
3. Signal processing in the Wavelet Domain.

3.1. Manipulation of the intrinsic pitch.
Intrinsic pitch information in the fine structure of the spectrum can be reconstructed from the wavelet coefficients corresponding to the zeroth and first dilation level. The periodicity in this fine structure is adjusted to target value using the “elimination-repetition” algorithm [2].

3.2. Computation of the optimal concatenation points.
Even if the phone boundaries could be detected perfectly during labelling, optimal coupling points may differ. Ideal coupling does not guarantee a maximum order of spectral similarity of the “PSOLA-bells” adjacent to the concatenation point. Naturally produced speech itself is highly nonstationary in the transitions between phones.
In the ideal coupling process, the right hand side unit (unit B in Figure 4) would have a “history” with exactly the spectral envelope as the spectral envelope of the lefthand side unit (unit A in Figure 4). Analogously the “PSOLA-bells” taken from the original context of the lefthand side unit following after the concatenation point should be as similar as possible to the right hand side unit.

For to quantise the mismatch in the context of the pitches to be synthesised we use the sum of the distances of the spectral envelopes of the adjacent frames (see figure 4).
In computation we use the squared distances of the mid range dilation wavelet coefficients of the “PSOLA-bells” within a window size of two.
In search of the optimal concatenation point we compute this mismatch for unit boundaries floating in a range of +/-1. (In Figure 4 the right side bound of unit A and the left hand side bound of unit B would be varied.) The combination with the smallest computed distance is used in concatenation.

3.3. Adjusting the spectral shape.
After optimising the concatenation points the articulatory tracks that produced the adjacent units are well synchronised but not identical.
For to avoid unnatural glitches in spectral shape at the concatenation points we use a fuzzy change between the adjacent articulatory tracks. The spectral shapes of the pitches in a +/-2 neighbourhood of the concatenation point are a weighted sum of the shapes in the “PSOLA-bells” of both articulatory tracks. The weights are linear with the distance form the concatenation point.

Following the multiresolution approach each complex wavelet coefficient in the considered pitches is adjusted separately.

Within the multilingual Siemens TTS system “Papageno” automatically generated phone-size units are concatenated. Automatical segmention of speech still is less accurate than manual tuning of units. Concatenation in the nonstationary transition between phones demands a dynamic adjustment of the units at the boundaries. So finding the optimal coupling point and smoothing spectral irregularities is more important than within conventional diphone concatenation.
The analysis stage includes windowing, FFT, and magnitudes during data preparation. Storing the synthesis inventory in the spectral domain, enables highly efficient coding techniques [11]. Unit (boundaries) are expanded by a delta of 2 pitches. This overlap is used in search of the optimal coupling points (see 3.2).

When performing synthesis we switch to the wavelet domain. After signal manipulation in the the wavelet domain we proceed the other way round via the frequency domain back to the time domain. Run time stages are:
• fast wavelet transform (FWT),
• choosing a suitable coupling point,
• smoothing irregularities in the spectral shape, represented by the midrange wavelet coefficients,
• inverse FWT,
• adjusting the phase of the frequency bands to the altered magnitude [2],
• inverse FFT,
• finally in time domain the hamming windowed signals are overlapped and added at the target pitchrate. [1]
5. Implementation in a dyadic filterbank structure.

The ideal wavelet transform discussed in section 2 can be realised very efficiently in a dyadic filterbank structure [8]. Decomposing the signal in complex wavelets we need one filter (F1) for real and one (F2) for the imaginary part of the wavelet. The left group of filters in figure 6 calculates the complex wavelet coefficients and the real lowpass residuum (F0) at a dilation level a. The right group of filters in figure 6 H0, H1, H2 realise the inner sum in formula (1). They are reconstructing the signal from the coefficients and the lowpass residuum at every dilation level. The downsampling represents the shift in equations (1) and (2).

![Figure 6: Analysis synthesis filterbank for a complex wavelet at one dilation level a.](image)

The fast wavelet transform recursively decomposes a signal s(f) to all the dilates of the wavelet function until the lowpass residuum $\xi(f)$ reduces to a constant. This can be realised very efficiently by a tree structure recursion of the filterbank of figure 6 as shown in Figure 7. [9]

The complexity of the FWT is linear in the length of the analysis and synthesis filters. [9]. Using 7-, 5-, 5- taps the filter coefficients of H0, H1, H2 are optimised to approximate the ideal wavelet function (3) (4) with maximal regularity. Figure 7 shows the convergence of the real and the complex wavelets at the fifth recursion level together with their shared scaling function.

![Figure 7: Synthesis Filters H0, H1, H2 after their fifth recursion in a tree structured filter bank.](image)

6. Discussion and conclusions.

The proposed techniques allow a completely non parametric analysis, and adjustment of complex spectral shapes, preserving a high degree of naturality in the synthetic speech. The representation as a linear combination of the perceptual relevant patterns offers a unified framework for spectral analysis and manipulation.

Informal listening test have shown that the algorithms for optimising the concatenation point and adjusting the spectral shape performed in the wavelet domain significantly improved the overall quality and naturalness of synthetic speech.

Low computational cost of fast algorithms in FWT and FFT make this methods suitable for real time processing.

There remain some improvements for future work. The choosen wavelet representation of the spectrum does not exploit the logarithmic frequency resolution of the human basilar membrane. Psychoacoustical characteristics as temporal and simultaneous masking should be made use of, to avoid perceptually irrelevant signal manipulation at the concatenation points.

7. References.

[1] C. Hamon, E. Moulines, and F. Charpentier, “A diphones synthesis system based on time domain modification of speech” ICASSP 89, Glasgow
[9] Strang Gilbert and Truong Nguyen, ”Wavelets and Filter Banks” Wellesly Cambridge Press, 96