SPECTRAL SUBTRACTION USING A NON-CRITICALLY DECIMATED DISCRETE WAVELET TRANSFORM

Andreas Engelsberg and Thomas Gülzow

Institute for Network and System Theory,
Technical Department, Kiel University,
Kaiserstrasse 2, D-24143 Kiel / Germany,
E-mail: ae@techfak.uni-kiel.de and tg@techfak.uni-kiel.de

ABSTRACT

The method of spectral subtraction has become very popular in speech enhancement. It is performed by modifying the spectral amplitudes of the disturbed signal. The spectral analysis of the signal is usually done by a Discrete Fourier Transformation (DFT).

We propose a spectral transformation with nonuniform bandwidth to take into account the characteristics of the human ear. The spectral analysis and synthesis is performed by a non-critically decimated discrete wavelet transform. Critical subsampling is not performed to avoid errors due to aliasing.

A significant drawback of spectral-subtraction methods are tonal residual noises in speech pauses with unnatural sound. The application of the proposed wavelet transform results in reduced residual noise with subjectively more comfortable sound.

1. INTRODUCTION

This publication deals with the enhancement of speech signals. A popular method to reduce additive noise of speech signals is spectral subtraction [1]. The basic idea of spectral subtraction is to estimate the magnitude of the noise-free spectrum by subtracting a mean magnitude of the noise spectrum from the disturbed spectrum.

In practice the required spectral analysis and synthesis is usually performed by a DFT and its inverse [1] or by analysis and synthesis filterbanks, for example polyphase filterbanks [2]. All systems have in common that they perform a uniformly spaced division of the frequency domain. The spectral analysis of the human ear can be modeled as a nonuniform filterbank with bark-scaled frequency bands [6]. This model was successfully applied to, e.g., speech recognition and coding systems.

We propose a non-critically decimated discrete wavelet filterbank in conjunction with spectral-subtraction methods. The wavelet filterbank is designed to approximate the frequency analysis of the human ear.

To avoid signal distortion due to spectral alias by modifying the spectral magnitude of the noisy signal, the decimation in each frequency band remains above the critical decimation by a factor of 2. The wavelet filterbank is based on a scaled Morlet wavelet and additional subfilters within each octave.

The paper is organized as follows:
First a brief review of speech enhancement using spectral subtraction is given. Then the analysis and synthesis stages of the wavelet filterbank are described in detail. In the next chapter experimental results are presented and a discussion of the results is given.

2. SPECTRAL SUBTRACTION

The spectral-subtraction method can be applied to noisy speech signals with additive noise

$$x(k) = s(k) + n(k) \quad , \tag{1}$$

where $x(k)$ denotes the noisy signal, $s(k)$ the speech signal and $n(k)$ the noise. The Fourier transformation is denoted as

$$\mathcal{F}\{x(k)\} = X(e^{j\Omega}) = |X(e^{j\Omega})| \cdot e^{j\varphi_x(\Omega)} \quad . \tag{2}$$

The basic idea is to subtract an estimated mean spectral magnitude of noise $|\overline{N(e^{j\Omega})}|$ from the spectral magnitude $|X(e^{j\Omega})|$ of the noisy speech signal. An estimation of the noise-free speech spectrum results as

$$\hat{S}(e^{j\Omega}) = \left(\max\left(|X(e^{j\Omega})| - |\overline{N(e^{j\Omega})}|, 0\right)\right) \cdot e^{j\varphi_x(\Omega)} \quad . \tag{3}$$

Note that only the magnitude of $X(e^{j\Omega})$ is modified, but the phase of the disturbed speech signal is preserved. The human ear is relatively insensitive to disturbances of the phase, so the exclusive modification of the magnitude is justified [2].

Negative values of $|X(e^{j\Omega})| - |\overline{N(e^{j\Omega})}|$ are estimation errors and therefore forced to zero. The mean magnitude of the noise spectrum is assumed to be estimated e.g. during speech pauses. This requires a separate speech-pause detector. An alternative is spectral-minima tracking [3].

Equation (3) may be interpreted in terms of spectral weighting of the noisy speech signal. This inter-

pretation leads to

$$\hat{S}(e^{j\Omega}) = H(e^{j\Omega}) \cdot X(e^{j\Omega}) \qquad (4)$$

where

$$H(e^{j\Omega}) = \max\left(\frac{|X(e^{j\Omega})| - |\overline{N(e^{j\Omega})}|}{|X(e^{j\Omega})|}\,,\,0\right)$$

$$= \max\left(1 - \frac{|\overline{N(e^{j\Omega})}|}{|X(e^{j\Omega})|}\,,\,0\right) \quad . \qquad (5)$$

The spectral weights $H(e^{j\Omega})$ are signal dependent and real. In practical realizations equation (4) has to be dicretized. This may be done by various kinds of spectral transformations or filterbanks.

Unfortunately residual tonal noises with unnatural sound remains especially in speech pauses. This is a major drawback of the spectral-subtraction method. More sophisticated spectral-subtracting rules were developed in the past [5, 4] to suppress the tonal noises.

Our experiments were developed with the basic system described above due to BOLL [1] and the subtraction rule due to EPHRAIM AND MALAH [5].

3. WAVELET TRANSFORM

The continuous wavelet transform (CWT) of a signal $x(t)$ is defined as

$$\mathcal{W}_x^{\psi}(b,a) = |a|^{-\frac{1}{2}} \int_{-\infty}^{+\infty} x(t)\psi^*\left(\frac{t-b}{a}\right) dt \quad . \qquad (6)$$

All basis functions $\psi_{b,a}(t) = |a|^{-\frac{1}{2}}\psi\left(\frac{t-b}{a}\right)$ are obtained by dilation or contraction from one single prototype wavelet $\psi(t)$. Large values of a cause $\psi_{b,a}(t)$ to become a lower-frequency and dilated version of $\psi(t)$. For small a values, the function $\psi_{b,a}(t)$ becomes a contracted version of $\psi(t)$ with higher frequency components. As a consequence, the resolution in the time-frequency plane is not constant. For high frequencies the resolution of the wavelet transform is sharp in time but poor in frequency, while for small frequencies the resolution is sharp in frequency and poor in time.

In the frequency domain the wavelet transform can be interpreted as a filterbank with bandpasses whose bandwidths $\Delta\omega_i$ increase monotonously with the center frequency ω_{0_i}. It can be shown that the relative bandwidth $Q = \frac{\Delta\omega_i}{\omega_{0_i}}$ is independent from the parameter a, so the wavelet-transform is called 'constant-Q' analysis. This is very similar to the frequency analysis of the human ear.

The digital realization of (6) requires the discretization of the parameters a and b. Usually the parameter a and b are chosen to be on a dyadic grid. Then a is a power of 2 and b is dependent on a, so that $a_m = 2^m, b_{mn} = a_m nT$, where $m, n \in \mathbb{Z}$. On this basis the wavelet transform in application to a discrete signal $x(k)$ becomes

$$w_x^{\psi}(2^m n, 2^m) = 2^{-\frac{m}{2}} \sum_k x(k)\psi^*\left(2^{-m}k - n\right) \quad , \qquad (7)$$

and realizes an octave analysis with different sampling rates in each octave.

To increase the resolution in frequency by a factor M, it is possible to use M dyadic wavelet analyses (voicing) each with the scaled prototype wavelet

$$\psi^j(k) = 2^{-\frac{j}{2M}}\psi(2^{-\frac{j}{M}}k), \quad j = 0, ..., (M-1) \quad . \qquad (8)$$

The non-critically decimated wavelet filterbank is based on the Á-Trous algorithm [7] and is one realization of (7). It can be shown that the filterbank structure in figure 1 approximates the discrete wavelet

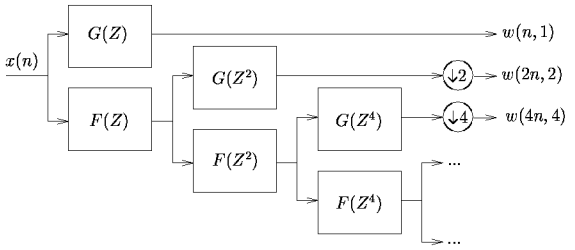

Figure 1: realization of the Á-Trous algorithm.

transform of (7) with non critical subsampling. $G(z)$ is the Z-transform of the prototype wavelet and $F(Z)$ is the Z-transform of an interpolation filter. While using the resolution of identity for multirate systems the decimation of 2^l in the l-th octave can be done more effectly and leads to the realized wavelet-filterbank structure shown in figure 2.

The bandpass filters $g^i(n)$, $i = 0, .., (M-1)$, are the prototype wavelets for each dyadic wavelet analysis. The decimation of 2^l in the l-th octave as shown in figure 2 allows the use of the same filter within each octave. The function of the lowpass filter $f_a(n)$ may be interpreted as an antialiasing-filter.

The synthesis filterbank interpolates the subbands to the next higher sampling rate and adds the result to the output of the next octave, taking care of the correct delay as produced in the analysis part.

4. EXPERIMENTS

The experiments were carried out using the above described wavelet filterbank with 7 octaves ($p=6$ in figure 2). We chose two different values for the number of voices, to investigate the influence of the number of subbands on the enhancement system. The choice of 10 voices per octave ($M=10$) leads to 70 channels

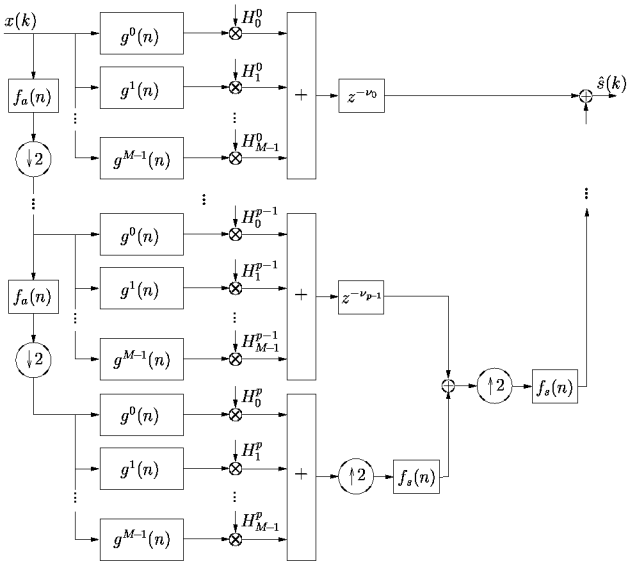

Figure 2: structure of the realized wavelet filterbank.

and 35 subbands results from applying 5 voices per octave ($M=5$).

The prototype filter is the sampled complex Morlet wavelet

$$\psi(t) = e^{j\omega_0 t} e^{-\frac{\beta^2 t^2}{2}} \tag{9}$$

[8] with 101 coefficients for the wavelet-filterbank with 70 subbands and 49 coefficients for the filterbank with 35 subbands. This choice yields a sufficient attenuation to the neighboring subband. The interpolation was implemented by FIR filters of length 71. The resulting frequency resolutions of the filterbanks are shown in figure 3. For the sake of clearness the plots are only shown for the first three octaves.

We examined the enhancement system with the spectral-subtracting rule due to BOLL [1] and the subtraction rule due to EPHRAIM AND MALAH [5]. The basic difference between uniform and non-uniform spectral analysis is well understandable applying the wavelet filterbanks to the basic spectral-subtraction system of BOLL, where a high amount of residual noises occurs. On the other hand we investigate, if the non-uniform spectral analysis gives improvements to more sophisticated spectral-subtraction rules with less residual noises.

In figure 4 four spectrograms are plotted. The results of the spectral-subtraction rule due to BOLL with different filterbanks are shown. The speech signal was recorded in a running car; so the noise was produced by the engine, the wheels and wind-turbulences. The sampling frequency was 11.025 kHz.

The frequency representation of the noisy speech signal is visualized in the first spectrogram. Most of the noise energy is located in the low-frequency area.

The data of the second spectrogram were produced using a polyphase filterbank with 256 chan-

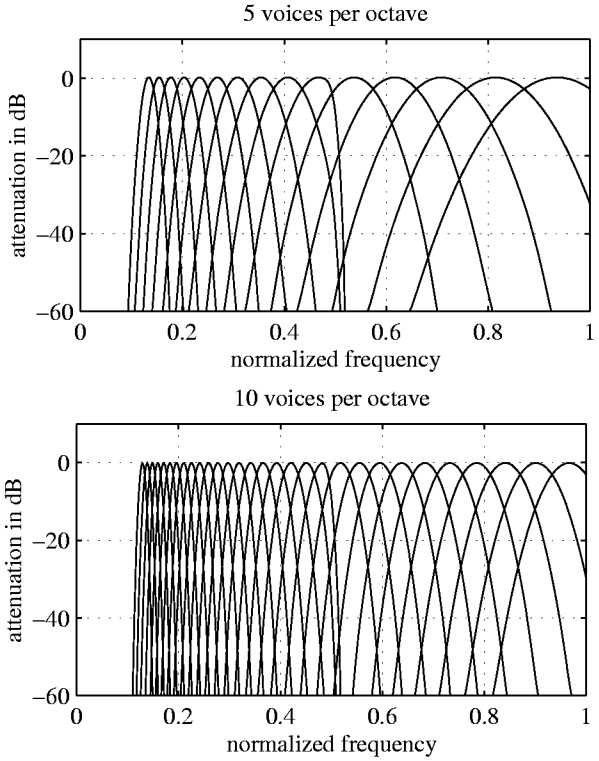

Figure 3: frequency resolutions for the first three octaves.

nels. The length of the prototype filter was chosen to 1024 samples and the decimation was performed with a factor 2 above critical subsampling. Especially in speech pauses randomly spaced spectral peaks remain, which produce tonal residual noises.

The third and fourth spectrogram show the results using wavelet filterbanks with 70 and 35 channels as spectral transformation.

The main differences appear in the structure of the tonal residuals in speech pauses. The spectral peaks in the second spectrogram are of uniform bandwidths, while the bandwidths of the spectral peaks in the third and fourth spectrogram increase to higher frequency but the duration in time decreases. Note that the amount of residual spectral peaks in higher frequency areas is significantly reduced using the wavelet filterbanks. The usage of fewer channels leads to fewer residual noises towards higher frequencies. But reducing the number of channels below 35 becomes problematic, because the filterbank is not able to separate the areas between the pitch frequencies. No noise reduction in these areas can be performed.

In informal listening tests the residual noise produced by the wavelet-based enhancement systems was judged to be more pleasant. The 70 channel wavelet filterbank was preferred to the 35 channel filterbank, because the amount of noise reduction to very low frequencies was higher.

The application of the spectral-subtraction rule of EPHRAIM AND MALAH leads to equivalent subjec-

tive results. Because the amount of residual noises is lower then in the case of BOLL's procedure, the advantages of the non-uniform spectral analysis can be used to adjust the parameters to achieve less distortion of the speech signal.

5. CONCLUSION

A wavelet-based spectral-subtraction system is proposed. Informal listening tests showed a subjective preference of filterbanks with nonuniform bandwidths for spectral-subtraction systems. In experimental investigations a choice of 70 channels for the proposed wavelet filterbank was found to be appropriate.

6. REFERENCES

[1] Boll, Steven F. "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 27, pp 113-120, April 1979

[2] Vary, Peter "Noise Suppression by Spectral Magnitude Estimation – Mechanism and Theoretical Limits –", EURASIP Signal Processing, Vol. 8, No. 4, pp 387-400, July 1985

[3] Martin, Rainer "Spectral Subtraction Based on Minimum Statistics", Proceedings of the EUSIPCO, Edinburgh, pp 1182-1185, 1994

[4] Cappé, Olivier "Elimination of the Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor", IEEE Transactions on Speech and Audio Processing, Vol. 2, pp 345-349, April 1994

[5] Ephraim, Y and Malah, D "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator", IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 32, pp 1109-1121, December 1984

[6] Zwicker, E. "Psychoakustik", Springer Verlag, 1982

[7] Shensa, M.J.: "'The Discrete Wavelet Transform: Wedding the Á Trous and Mallat Algorithms"', *IEEE Transactions on Signal Processing*, Vol. 40, No. 10, pp. 2464-2482, October 1992.

[8] Daubechies, I.: *Ten Lectures on Wavelets*, Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, 1992.

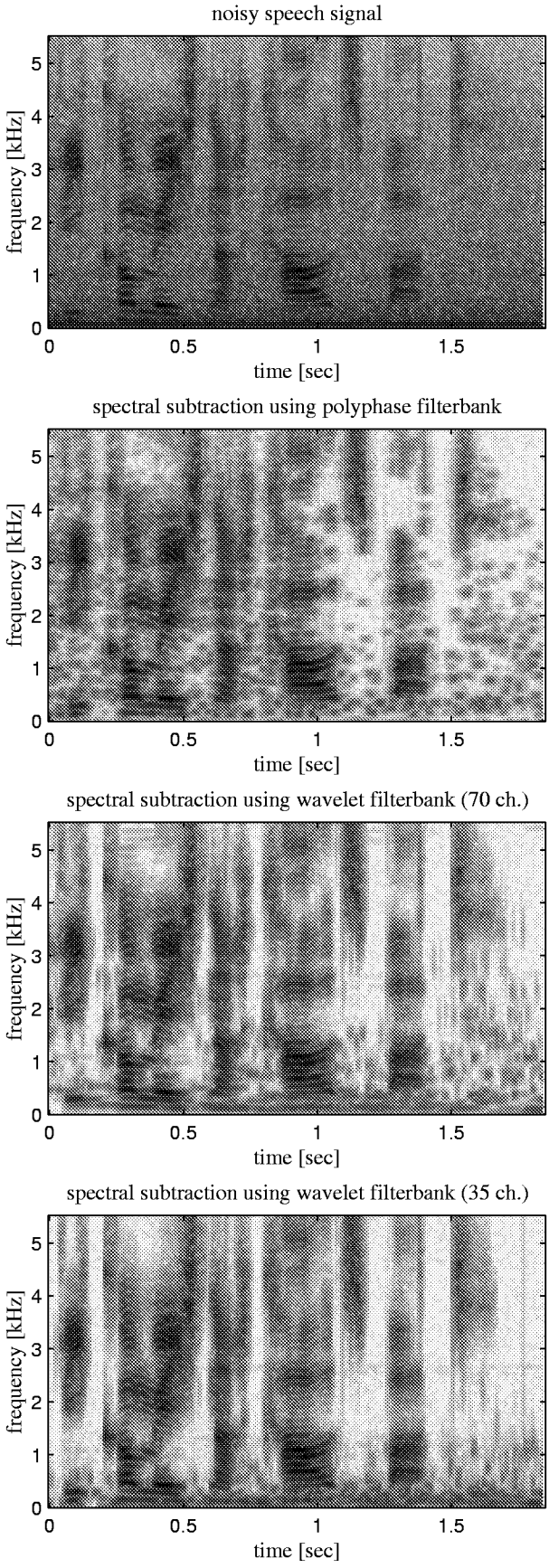

Figure 4: comparison of spectrograms.