SPEECH ENHANCEMENT FOR MOBILE TELEPHONY BASED ON NON-UNIFORMLY SPACED FREQUENCY RESOLUTION

Pia Dreiseitel and Henning Puder
Signal Theory, Darmstadt University of Technology
Merckstr. 25, D–64283 Darmstadt, Germany
{dreiseit,hpuder}@nesi.tu-darmstadt.de

ABSTRACT

In this paper, we present a speech enhancement method based on spectral subtraction in non-uniformly spaced sub-bands. The main advantage of using different frequency resolutions for the various bands is the perception property of the human ear, which is able to separate low frequencies more precisely than high frequencies. The parameters of the noise reduction algorithm are chosen appropriately to the respective signal to noise ratio, which yields a performance superior to an uniform frequency resolution with the same number of subbands. A two-stage cascaded filter-bank is used for the decomposition of the signal.

1 INTRODUCTION

When a hands-free telephone system is applied in a car environment, it is desirable to have an enhancement of the incoming signal consisting of speech and background noise of a fairly low signal to noise ratio. Noise reduction, which can be used here, has been discussed for several years but there is still no solution available which satisfies the noise reduction problem and leaves the speech without distortion. Spectral subtraction rules are commonly applied for noise reduction of a single channel speech signal [5, 2]. The signal is split up into its spectral components by either using a short-time Fourier Transform (STFT) or by using a polyphase filter-bank. For efficient noise reduction of speech signals, it is necessary to use a fine frequency resolution of at least 40 Hz, especially in the lower frequency bands. Since speech signals cannot be treated as stationary signals, the analysis window may be to long to meet the stationarity restrictions of speech. This leads to a trade off concerning the window length between frequency resolution and stationarity properties of the input signal. Speech signals are usually assumed to be stationary for about 20–30 ms. A common window size for the STFT analysis is therefore about 256 samples (being the equivalent of 32 ms) when dealing with signals sampled at 8 kHz. This leads to a frequency resolution of about 30 Hz which may be sensitive enough for achieving good noise reduction results in the lower bands. A non-uniformly spaced frequency resolution outperforms the equally spaced spectral analysis, because it allows to analyse the spectral components in an appropriate way. Taking short analysis windows for high frequency components and long ones for low frequency components leads to a high resolution in the lower bands and broad frequency bands in the upper part of the total frequency range. Since the human perception of audio signals also works in a logarithmic scale [7], this way of signal analysis seems appropriate for speech enhancement.

2 NON-UNIFORMLY SPACED FREQUENCY RESOLUTION

When dealing with a non-uniformly spaced frequency resolution, there are a number of possibilities for the decomposition. The wavelet transform has been discussed with respect to noise reduction of speech signals as well as a polyphase approach using all-pass filters instead of delay elements [3].

Wavelet decomposition

In recent years the wavelet decomposition became popular for all kinds of signal processing applications. However, for splitting an acoustic signal up into almost any number of sub-bands, this decomposition seems not particularly well suited. The wavelet decomposition, mostly implemented using a quadrature mirror filter structure, always bisects the signal at each stage. A relatively large number of stages is therefore required for a typical partitioning of the incoming signal (e.g. 8 stages). If subsampling is desired, there is only the possibility of the critical sampling rate. This also limits the performance because of high aliasing-terms introduced.

Modified filter-bank

Compared to the efficient implementation of polyphase filter-banks [1], the delay elements have to be replaced by all-pass filters to obtain the modified filter-bank with a non-uniform frequency resolution. This turns out to be very expensive in terms of computational cost.

Due to the phase distortions introduced by all-pass elements, an additional equalization has to be implemented. Therefore the total impulse response of the analysis and synthesis has to be determined and finally the signal has to be filtered by the inverse of the impulse response. This of course only guarantees a perfect performance when no further processing (like noise reduction) is implemented in between decomposition and synthesis.

Cascaded polyphase filter-bank

Our approach splits up the signal into sub-bands of different widths using a two stage cascaded polyphase filter-bank (s. Fig. 1). The signal is first split up into a moderate number of sub-bands, sub-sampled, and then split up once more using different frequency resolutions for the various sub-bands. The number of sub-bands is chosen with respect to the Bark scale [7] and the performance of the speech enhancement filter. The polyphase filter-banks applied for speech enhancement are perfect reconstruction filter-banks designed according to [6].

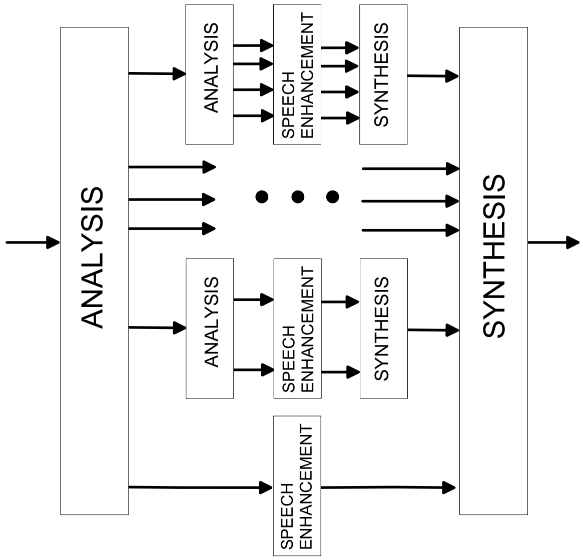

Figure 1: Cascaded filter-bank

When different frequency resolutions are used for the second stage of the cascaded filter-bank, one has to ensure that the delay introduced is equal for all sub-bands. Otherwise, an additional delay has to be added. Attention has to be paid on an equal gain of each sub-band decomposition.

3 SPECTRAL SUBTRACTION

For speech enhancement, a spectral subtraction rule is applied as follows:

$$\underline{G}_{opt}(k,n) = \begin{cases} 1 - \sqrt{\frac{\kappa \underline{N}_{PSD}(k,n)}{\underline{X}_{PSD}(k,n)}} & \underline{G}_{opt} > \beta_f \\ \beta_f & \textit{otherwise} \end{cases} \qquad (1)$$

where $\underline{N}_{PSD}(k,n)$ is the estimated power spectral density of the noise at time n and frequency k, $\underline{X}_{PSD}(k,n)$ the estimated power spectral density of the incoming signal comprising both noise and speech signal, β_f a spectral floor for limiting the maximum attenuation applied by the spectral subtraction and κ an overestimation factor. If no reference is available for the spectral behaviour of the noise signal, either voice activity detection or minimum statistics [4] have to be used to estimate $\underline{N}_{PSD}(k,n)$. Our approach uses voice activity detection:

$$\underline{N}_{PSD}(k,n) = \begin{cases} \lambda_{slow}\, \underline{N}_{PSD}(k,n-1) & \text{voice} \\ +(1-\lambda_{slow})\,|X(k,n)|^2 & \text{detected} \\ \underline{N}_{PSD}(k,n-1) & \text{otherwise} \end{cases} \qquad (2)$$

$$\underline{X}_{PSD}(k,n) = \lambda_{fast}\, \underline{X}_{PSD}(k,n-1) + (1-\lambda_{fast})\,|X(k,n)|^2 \qquad (3)$$

Since the sub-bands of the final decomposition do not have the same sampling rate, the weighting with G_{opt} cannot be done globally. A different speech enhancement filter is therefore derived for each set of sub-bands. However, this introduces new degrees of freedom for the constants used in equation (1).

Both $\underline{N}_{PSD}(k,n)$ and $\underline{X}_{PSD}(k,n)$ are estimated using recursively smoothed periodogramms, with λ_{fast} and λ_{slow} denoting the smoothing constants for fast and slow changing signals respectively. These constants are adapted to the respective sub-sampling rate.

4 OPTIMISED CHOICE OF PARAMETERS

In contrast to the one-step noise reduction procedure, the choice of parameters can be varied for every single set of sub-bands.

4.1 Sub-band frequency resolution

As mentioned before, the sub-band width for the lower frequency bands should not exceed about 40 Hz. This is also supported by the width of singular peaks of tonal speech components. In Fig. 2 and 3 the power spectral densities of the vowel "o" are shown, using different window sizes. It is obvious that for the window size 128 the peaks in the lower frequency bands are no longer separable, which worsens the speech enhancement capabilities.

For the highly varying high frequency components, a low resolution modelling only the average of the power spectral density is sufficient due to the noise like properties of the high frequency components.

4.2 Minimising distortions

When spectral subtraction methods are applied for noise reduction of speech signals, well known unnatural bubbling sounds, called *musical tones*, appear. The appearance of musical tones can be suppressed by a large overestimation factor κ and a high noise-floor β_f, which, limits the overall performance of the noise reduction system. It is therefore reasonable to detect the frequency bands where this effect is most disturbing. Psychoacoustic masking effects – noise components are not audible

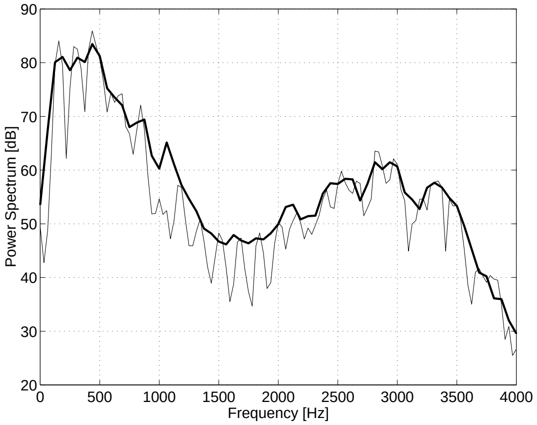

Figure 2: Power spectral density, window length 256 and 128(thick line), 256 samples correspond to 31 Hz, 128 to 62 Hz frequency resolution.

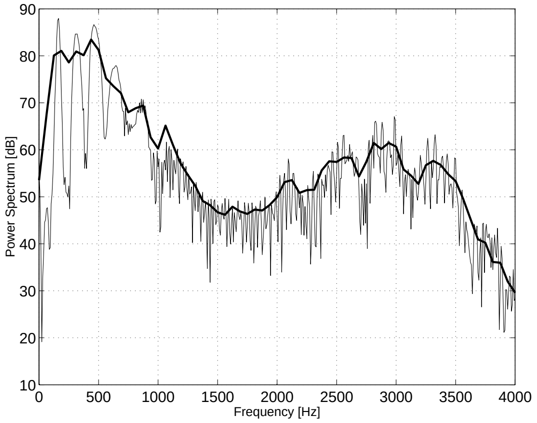

Figure 3: Power spectral density, window length 1024 and 128(thick line)

when superimposed by loud speech components in the same frequency range – make a high noise floor reasonable for frequency bands with relatively high signal to noise ratios.

The signal in this two-step approach is first split into 16 bands and then split again approximating the Bark scale. As the input signal is real, the first stage delivers symmetric sub-bands. Thus, one only has to process the lower half of the total bands.

As it is depicted in Tab. 1, both the spectral floor β_f and the overestimation factor κ are chosen appropriately to the respective signal to noise ratio. Car noise, in general, has most of its power in the lower frequency bands. The signal to noise ratio, therefore improves with increasing frequencies. The spectral floor and the overestimation factor can be smaller in case of a better signal to noise ratio.

Table 1: Overview on parameters for non-linear frequency decomposition

frequency band	resolution	No. of sub-bands	β_f	κ
0-250 Hz	16 Hz	32	0.10	2.0
250-750 Hz	16 Hz	32	0.10	2.0
750-1250 Hz	31 Hz	16	0.08	2.0
1250-1750 Hz	31 Hz	16	0.08	2.0
1750-2250 Hz	62 Hz	8	0.08	1.5
2250-2750 Hz	125 Hz	4	0.07	1.5
2750-3250 Hz	125 Hz	4	0.07	1.5
3250-3750 Hz	500 Hz	1	0.05	1.5
3750-4000 Hz	250 Hz	1	0.05	1.5

5 RESULTS

The results were achieved by simulations using speech signals superimposed by noise signals recorded in a car at the speed of 120 km/h at a signal to noise ratio of about 3 dB. To draw a comparison to a noise reduction based on uniformly spaced frequency resolution, the system proposed here is compared to a decomposition with the same number of frequency bands.

Due to the increased flexibility of the cascade polyphase structure, it was possible to emphasize the noise reduction in sub-bands with low signal to noise ratios and leave sub-bands with better signal to noise ratios almost untouched. This leads to a better performance compared to a one-step polyphase structure.

Especially for the lower frequency bands, the higher resolution proofed to generate a more natural sounding speech signal than the uniformly spaced frequency resolution with an equal number of sub-bands. However, special attention has to be paid on the design of the polyphase filter-banks when cascaded structures are used. One has to ensure, that the group delay is equal for all sub-bands.

6 COMPUTATIONAL LOAD

The noise reduction system proposed above needs a cascaded filter-bank for the decomposition of the incoming speech signal. Since there are efficient ways of implementing polyphase filter-banks, the numerical cost of the decomposition is not very high. The equally distributed decomposition with M_0 sub-bands needs

$$\Theta \quad = \quad \frac{2}{f_{u0}} \left(L_0 + 2\, M_0 \log_2(M_0) \right), \tag{4}$$

multiplications per sample, where L_0 denotes the length of the prototype low-pass window (we use $L_0 = 4M_0$) and f_{u0} the sub-sampling rate. The first term describes the multiplications of the signal with the prototype low-pass impulse response, the second corresponds to the complex multiplications of the required FFTs.

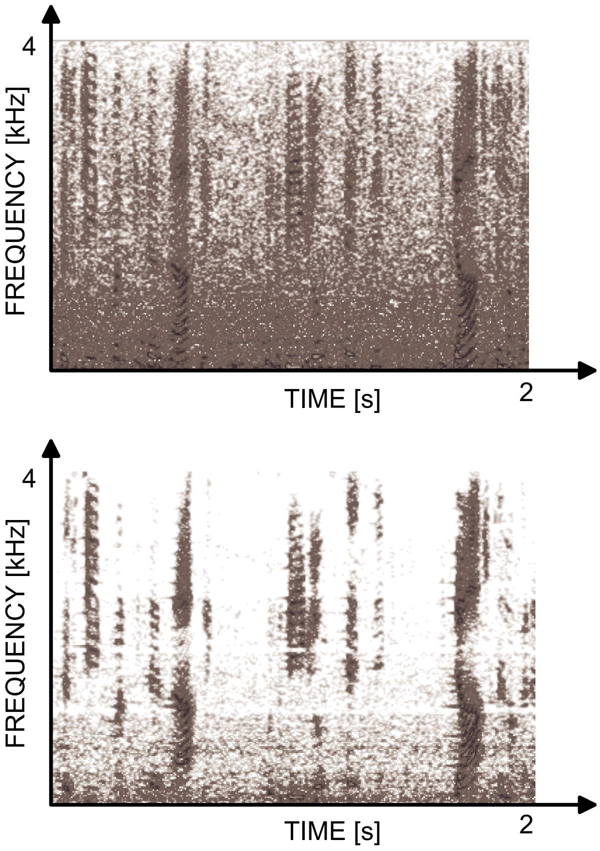

Figure 4: Spectrogramm of input and enhanced signal

For the cascaded filter-bank, the number of multiplications calculates from the sum of the first and second stage:

$$\begin{aligned}\Theta &= \frac{2}{f_{u1}}\left(L_1 + 2\,M_1 \log_2(M_1)\right) &(5)\\ &+ \frac{1}{f_{u1}} \sum_{i=0}^{M_1/2} \left[\frac{2}{f_{u2}(i)}\left[L_2(i) + 2\,M_2(i)\log_2(M_2(i))\right]\right],\end{aligned}$$

where $L_1, L_2(i)$ denote the lengths of the prototype low-pass windows of the first and second stage and $f_{u1}, f_{u2}(i)$ the respective sub-sampling rates. For the exemplary setting of Tab. 1, 87 multiplications are required per sample.

In contrast to this, there are no computationally efficient implementation for both the all-pass filter-bank and wavelet transforms. For a comparable non-uniform all-pass filter-bank, 256 additional multiplications per sample are necessary, since all-pass filters have to be calculated in the full sampling rate [3]. Wavelet transforms require

$$\Theta = 2\,M_w \sum_{i=0}^{N_s} 2^{-i} \approx\, 4\,M_w \qquad (6)$$

multiplications, with M_w being the length of the wavelet and N_s the number of stages. For wavelets longer than 22 taps, this decomposition affords more computational load than a cascaded filter-bank.

The delay introduced by the cascaded filter-bank, however, is not negligible. It consists of the window length of the first stage and additionally the largest window length of the second stage multiplied by the subsampling factor of the first stage. This is comparable to the delay of a uniformly spaced frequency decomposition having the highest resolution of the cascaded filter-bank for all bands.

7 CONCLUSIONS

We have presented a spectral subtraction for noise reduction based on a non-uniform frequency decomposition. This is realized by a two-stage cascaded filter-bank. Simulations have shown that this way of dealing with speech signals is superior to the uniformly spaced frequency resolution using the same number of sub-bands, especially for the lower frequency components. The enhanced speech signal keeps more components of the lower sub-bands and therefore sounds richer. The typical distortions occurring in spectral subtraction methods could be widely attenuated. In comparison to other proposals for non-uniform filter-banks, this approach has the advantage of lower computational requirements.

References

[1] R. Crochiere and L. Rabiner. *Multirate Digital Signal Processing.* Prentice Hall, Englewood Cliffs, 1983.

[2] P. Dreiseitel and H. Puder. Noise reduction and improved echo cancelation. In *Proceedings of the IWAENC 1997*, September 1997.

[3] T. Gülzow, A. Engelsberg, and U. Heute. Comparison of a discrete wavelet transform and a nonuniform polyphase filterbank applied to spectral subtraction speech enhancement. *EURASIP Signal Processing*, 64(1):5–19, January 1998.

[4] R. Martin. Spectral subtraction based on minimum statistics. In *Proceedings of the EUSIPCO 1994*, pages 1182–1185, Edinburgh, September 1994.

[5] P. Vary. Noise suppression by spectral magnitude estimation – mechanism and theoretical limits. *EURASIP Signal Processing*, 8(4):387–400, July 1985.

[6] G. Wackersreuther. On the design of filters for ideal qmf and polyphase filter banks. *AEÜ*, 39(2):123–130, 1985.

[7] E. Zwicker and R. Feldkeller. *Das Ohr als Nachrichtenempfänger.* Hirzel Verlag, Stuttgart, Germany, 1 edition, 1967.