Near End Listening Enhancement by Means of Warped Low Delay Filter-Banks

Bastian Sauert, Heinrich W. Löllmann, and Peter Vary

Institute of Communication Systems and Data Processing (**ind**), RWTH Aachen University, 52056 Aachen
E-Mail: {sauert,loellmann,vary}@ind.rwth-aachen.de
Web: www.ind.rwth-aachen.de

Abstract

The concept of near end listening enhancement allows to improve the speech intelligibility of telecommunication devices in the presence of ambient background noise. It raises adaptively the average speech spectrum of the received signal from the far end speaker over the average noise spectrum of the near end background noise, which leads to reduced listening efforts for the near end listener.

In this paper, we will propose an improved algorithm for near end listening enhancement which uses a non-uniform filter-bank with Bark-scaled frequency bands. The employed filter-bank has a significantly lower signal delay than commonly used non-uniform analysis-synthesis filter-banks. This allows to achieve a comparable speech intelligibility with low latency, which is of interest, e. g., for applications in mobile phones.

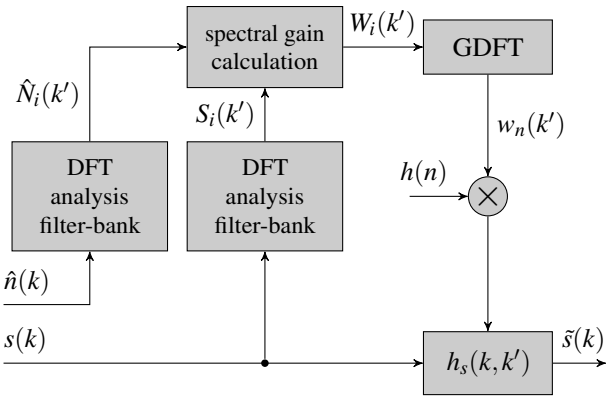


Figure 2: New system for near end listening enhancement with low signal delay.

1 Introduction

Mobile communication is often conducted in the presence of acoustical background noise such as traffic or babble noise. This leads to the problem that the *near end* listener, if located in a noisy environment, perceives a mixture of the clean *far end* speech and the acoustical background noise from the *near end* and thus experiences a reduced speech intelligibility.

For the problem of near end listening enhancement, as opposed to the problem of noise reduction, the noise signal cannot be influenced because the person is located in a noisy environment and the noise reaches the ear with hardly any possibility to intercept. Therefore, a reasonable approach to improve intelligibility by digital processing is to manipulate the *far end* speech signal in dependence of the *near end* background noise as depicted in Figure 1.

In [1], we have proposed a time and frequency dependent approach which tries to amplify the far end speech signal in order to reestablish a certain level difference between the average speech spectrum and the measured noise spectrum, i. e., to recover a target signal-to-noise ratio (SNR). This algorithm uses a common discrete Fourier transform (DFT) analysis-synthesis filter-bank (AS FB), implemented by the overlap-add method.

However, such a uniform DFT filter-bank does not account for the non-uniform frequency resolution of the human ear, which can be modeled by the Bark frequency bands, cf. [2]. A very good approximation of the Bark bands can be achieved by means of an allpass transformed filter-bank [3]. Such *frequency warped* filter-banks are beneficial for speech and audio processing as, for example, noise suppression systems, e. g., [4]. A drawback of warped AS FBs is their higher computational complexity and signal delay in comparison to a uniform AS FBs, which often precludes their use for applications such as mobile communication devices. An approach to realize a warped filter-bank with significantly lower complexity and signal delay than common warped AS FBs is given by the concept of the *filter-bank equalizer* (FBE) [5, 6]. By this, it is possible to exploit the benefits of a Bark-scaled frequency resolution without causing a high delay.

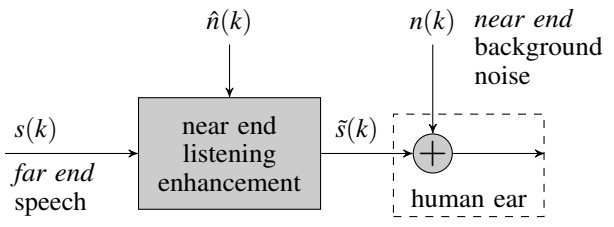


Figure 1: Principle of near end listening enhancement.

In this contribution, we will show how the use of the warped FBE for near end listening enhancement yields a comparable speech intelligibility with lower signal delay than by means of a uniform DFT AS FB. In [5], the concept of the *auto-regressive low delay filter* (AR LDF) is also introduced, which accomplishes a signal delay of only a few samples. The application for near end listening enhancement yields a high improvement of the speech intelligibility with very low signal delay, which is of interest for systems with tight latency constraints.

2 New Enhancement System

An overview of the new system for near end listening enhancement by means of the FBE is depicted in Figure 2. In essence, a time-domain filtering is performed with coefficients adapted in the frequency-domain. A description of the single processing blocks is given in the following.

2.1 Concept of Uniform FBE

The (clean) far end speech signal $s(k)$ and the near end noise estimate $\hat{n}(k)$ are split into M subband signals $S_i(k')$

and $\hat{N}_i(k')$ by means of a DFT analysis filter-bank with downsampling

$$S_i(k') = \sum_{n=0}^{L} s(k'-n)\cdot h(n)\cdot \exp\left\{-\mathrm{j}\tfrac{2\pi}{M} i\cdot n\right\}, \quad (1)$$

$$\hat{N}_i(k') = \sum_{n=0}^{L} \hat{n}(k'-n)\cdot h(n)\cdot \exp\left\{-\mathrm{j}\tfrac{2\pi}{M} i\cdot n\right\}, \quad (2)$$

$$i \in \{0,1,\ldots,M-1\}.$$

The subsampled time index is given by $k' = \lfloor k/R \rfloor \cdot R$ where R marks the downsampling rate. The real impulse response of the prototype filter of length $L+1$ is denoted by $h(n)$.

The subband signals $S_i(k')$ and $\hat{N}_i(k')$ are used to calculate the spectral gains $W_i(k')$ as described later in Section 2.4. The enhanced speech signal $\tilde{s}(k)$ is obtained by filtering the far end speech signal $s(k)$:

$$\tilde{s}(k) = \sum_{n=0}^{L} s(k-n)\cdot h_s(n,k'). \quad (3)$$

The time-varying coefficients of this filter are obtained from the spectral weights according to

$$h_s(n,k') = h(n)\cdot \sum_{i=0}^{M-1} W_i(k')\cdot \exp\left\{-\mathrm{j}\tfrac{2\pi}{M} i(n-n_0)\right\} \quad (4)$$

with $n \in \{0,1,\ldots,L\}$ and $n_0 = L/2$. Hence, the filter coefficients are obtained by a generalized discrete Fourier transform (GDFT) of the spectral weights $W_i(k')$.[1]

It should be noted that only the concept of the FBE is described here. In practice, the FBE can be efficiently implemented by means of a polyphase network with DFT and GDFT calculated efficiently by the fast Fourier transform (FFT). Such aspects are treated in [5, 6] in more detail.

2.2 Non-Uniform FBE

The FBE with non-uniform time-frequency resolution is designed by means of an allpass transformation. In the process, the delay elements of the discrete filters are replaced by (causal) allpass filters of first order

$$\mathrm{z}^{-1} \rightarrow H_A(\mathrm{z}) = \frac{\mathrm{z}^{-1}-a}{1-a\,\mathrm{z}^{-1}}\,;\quad -1<a<1\,. \quad (5)$$

The corresponding frequency response reads

$$H_A\left(\mathrm{e}^{\mathrm{j}\Omega}\right) = \mathrm{e}^{-\mathrm{j}\,\varphi_a(\Omega)} \quad (6)$$

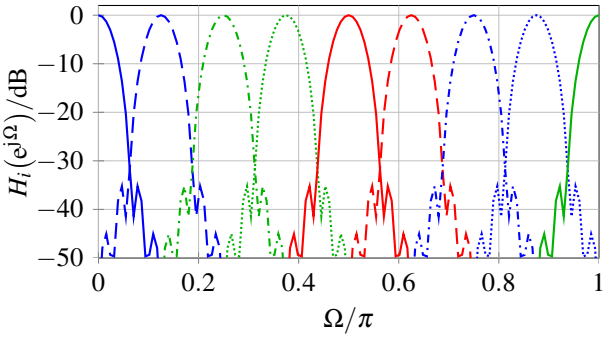
with

$$\varphi_a(\Omega) = 2\arctan\left(\frac{\sin(\Omega)}{\cos(\Omega)-a}\right) - \Omega\,. \quad (7)$$

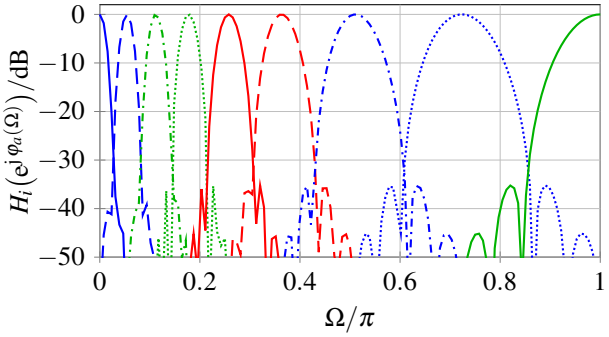
Due to this allpass transformation, the frequency responses of the uniform subband filters

$$H_i\left(\mathrm{z}=\mathrm{e}^{\mathrm{j}\Omega}\right) = \sum_{n=0}^{L} h(n)\cdot \mathrm{e}^{-\mathrm{j}\frac{2\pi}{M} i\cdot n}\cdot \mathrm{e}^{-\mathrm{j}\Omega n} \quad (8)$$

[1] The DFT instead of the GDFT analysis filter-bank is used here as the magnitude of the subband signals is only needed for the spectral gain adaptation according to Section 2.4.



(a) uniform subband filters



(b) warped subband filters

Figure 3: Magnitude responses of a uniform and warped DFT analysis filter-bank with $M = 16$ frequency bands and allpass coefficient $a = 0.4$.

are converted into warped subband filters with frequency responses

$$H_i\left(\mathrm{z}=\mathrm{e}^{\mathrm{j}\,\varphi_a(\Omega)}\right) = \sum_{n=0}^{L} h(n)\cdot \mathrm{e}^{-\mathrm{j}\frac{2\pi}{M} i\cdot n}\cdot \mathrm{e}^{-\mathrm{j}\,\varphi_a(\Omega) n}\,. \quad (9)$$

The effect of this frequency warping is illustrated in Figure 3. The allpass transformation achieves a variation of the bandwidths without changing filter properties such as stopband attenuation etc. An allpass pole of $a = 0.4$ yields a good approximation of the Bark frequency scale for the considered sampling rate of $f_s = 8\,\mathrm{kHz}$, cf. [3].

The allpass transformation changes not only the magnitude but also the phase response of the filters. This undesirable effect can be compensated by applying a phase equalizer to the output signal of the FBE [5, 7]. However, such phase equalization is not needed for a small filter degree L since the human auditory system is quite insensitive against phase modifications.

2.3 Low Delay Filter

The FBE achieves about half the algorithmic signal delay than the corresponding AS FB. However, a further reduction of the signal delay becomes necessary for applications with very tight signal delay constraints. A flexible approach to achieve this is to approximate the original time-domain filter of the FBE by a filter with lower degree. This concept, termed as low delay filter (LDF), can be accomplished by means of a moving-average (MA) filter approximation or an auto-regressive (AR) filter approximation [5]. In the following, the AR filter approximation is considered as this approach can achieve a delay of only a few samples.

For the *uniform* FBE, the time domain filter of (4) is

approximated by an AR filter with transfer function[2]

$$H_s(\mathrm{z}) \approx H_{AR}(\mathrm{z}) = \frac{b_0}{1 - \sum_{n=1}^{P} b_n \mathrm{z}^{-n}}. \qquad (10)$$

The $P+1$ AR filter coefficients b_n are determined from the $L+1$ coefficient $h_s(n)$ by means of the Yule-Walker equations, which can be efficiently solved by the Levinson-Durbin recursion [5]. The obtained AR filter is of minimum phase and therefore always stable.

This approach can also be applied for the *non-uniform* (warped) FBE, where the warped time-domain filter is approximated by a warped AR filter. The calculation of the AR filter coefficients is done as outlined before. Applying the allpass transformation of (5) to the AR filter of (10) in a straight-forward fashion leads to a filter with delay-free feedback loops which, however, can be solved by a modified filter structure [5, 6].

2.4 Spectral Weight Calculation

As described in [1], the time-varying gain factors $W_i(k')$ are chosen in a way that the ratio of the short-term power spectral density (PSD) of the amplified speech $\Phi_{\tilde{s}\tilde{s},i}(k')$ and the short-term PSD of the noise signal $\Phi_{nn,i}(k')$ should be greater than or equal to a target SNR ξ,

$$\frac{\Phi_{\tilde{s}\tilde{s},i}(k')}{\Phi_{nn,i}(k')} \geq \xi, \qquad (11)$$

with, e. g., $\xi \mathrel{\widehat{=}} 15\,\mathrm{dB}$. In order to assure that the speech signal is not attenuated in a noise-free environment, the gain factors may not be less than one so that

$$W_i'(k') = \max\left\{\sqrt{\xi \cdot \frac{\Phi_{nn,i}(k')}{\Phi_{ss,i}(k')}}, 1\right\}. \qquad (12)$$

The speech signal is weighted according to the spectral characteristics of the noise signal. However, the solution 'over-amplifies' low speech signal components since it tries to raise anything over the noise by the same amount ξ independent of the original signal strength. This effect can be reduced by limiting the gain $W_i'(k')$ to a maximum gain $W_{\max}$ with, e. g., $W_{\max} \mathrel{\widehat{=}} 30\,\mathrm{dB}$. It finally follows that

$$W_i(k') = \min\left\{\max\left\{\sqrt{\xi \frac{\Phi_{nn,i}(k')}{\Phi_{ss,i}(k')}}, 1\right\}, W_{\max}\right\}. \qquad (13)$$

The needed short-term PSD $\Phi_{ss,i}(k')$ is computed as the recursive average of the periodogram $|S_i(k')|^2$:

$$\Phi_{ss,i}(k') = \alpha_S \cdot \Phi_{ss,i}(k'-R) + (1-\alpha_S) \cdot \left|S_i(k')\right|^2 \qquad (14)$$

with time constant $0 < \alpha_S < 1$. The noise PSD $\Phi_{nn,i}(k')$ is obtained in the same manner with a (different) time constant α_N.

The choice of the time constants α_S and α_N is crucial for the performance of the algorithm. If they are too small, the amplified speech follows the noise too quickly and tends to lose its amplitude structure. If they are close to one, the system does not react to changing speech and noise signals. As discussed in [1], the values $\alpha_S = 0.996$ and $\alpha_N = 0.96$ turned out to be a good choice for the downsampling rate $R = 80$ and a prototype lowpass filter with $L+1 = 160$ coefficients.

[2]For the sake of simplicity, the time dependence of the filter is not considered here.

2.5 Adaption of Time Constants

The used warped filter-bank with Bark-scaled subbands has a non-uniform time-frequency resolution. For a positive allpass coefficient a, the subband filters at higher frequencies have higher bandwidths (lower frequency resolution) and a shorter impulse response (higher time resolution), and vice versa for the lower frequencies. This effect must be considered for the determination of the time constants α_S and α_N introduced in Section 2.4.

For a shorter impulse response, the time constants α_S and α_N must become greater in order to average over the same amount of samples. Therefore, the time constants are interpolated between the uniform case and the maximum value of one depending on the normalized bandwidth $\Delta\Omega_i$ of each subband:

$$\alpha_S'(i) = 1 - (1-\alpha_S) \cdot \frac{\Delta\Omega_{\text{uniform}}}{\Delta\Omega_i}. \qquad (15)$$

The uniform case of [1] uses 160 samples per DFT frame and, thus, has effectively 160 uniform subbands leading to a normalized bandwidth of $\Delta\Omega_{\text{uniform}} = \frac{2\pi}{160}$. The normalized bandwidth $\Delta\Omega_i$ of each Bark-scaled subband can be approximated by the distance of the normalized center frequencies of the subbands:

$$\Delta\Omega_i = \varphi_a(\Omega_i) - \varphi_a(\Omega_{i-1}) \quad \text{with} \quad \Omega_i = 2\tfrac{\pi}{M} i. \qquad (16)$$

The devised method is of course a heuristic approach which, however, yields good results in practice.

3 Results

The performance of the proposed algorithms was evaluated in terms of the Speech Intelligibility Index (SII) [8].

3.1 Speech Intelligibility Index

The SII is supposed to be correlated with the intelligibility of speech under a variety of adverse listening conditions. It is basically computed by adding the speech-to-noise ratio in each contributing frequency band weighted according to its contribution to speech intelligibility. According to [8], good communication systems have an SII of 0.75 or above, while poor communication systems have an SII below 0.45.

The SII was calculated with the so-called critical band procedure. In order to calculate the speech and noise spectrum level of each sound file, the spectrum level is averaged for half-overlapping Hann-windowed frames of 20 ms length. Thereby, an average speech spectrum level of the whole speech database was achieved which is comparable to the standard speech spectrum level for normal vocal effort specified in [8].

3.2 Simulation Results

In our evaluation, the SII was calculated for every speech file of the TIMIT database, in total 5.4 hours, disturbed by the *factory1* noise from the NOISEX-92 database for a sampling rate of 8 kHz. The mean values of the SIIs without processing and after processing with the proposed enhancement system, i. e., a warped FBE and warped AR LDF with $M = 64$ subbands, are depicted in Figure 4 for several signal-to-noise ratios.

For comparison the mean SII after processing with a uniform AS FB with $M = 256$ subbands as proposed in [1] is also plotted. As a second reference, each half-overlapping speech frame of 160 samples length is amplified with one

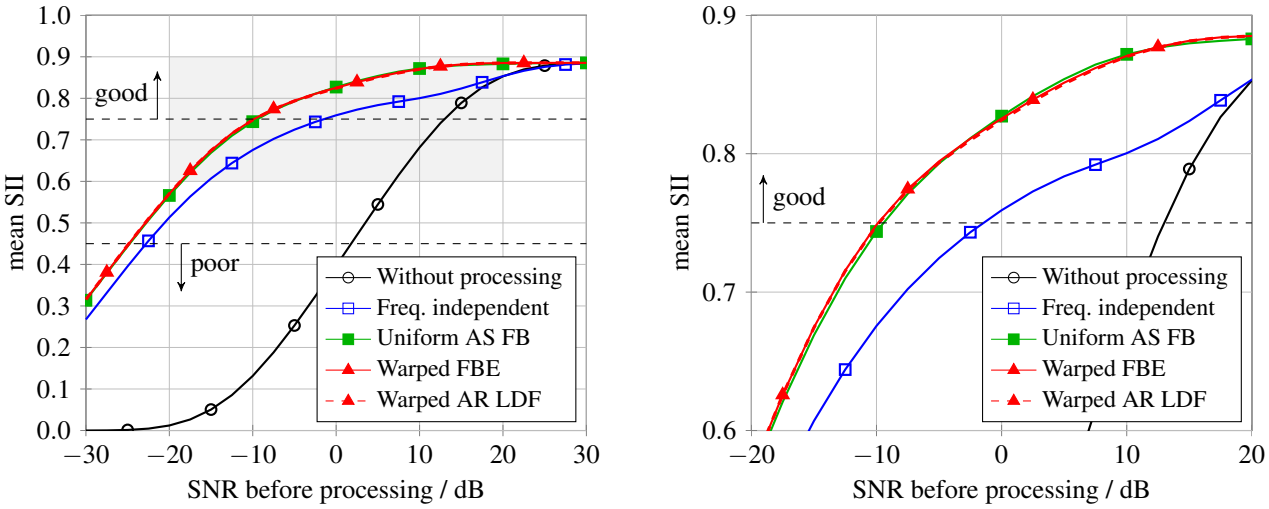


Figure 4: Comparison of mean SII after time-domain frequency independent processing as well as after frequency dependent processing using a uniform AS FB with $M = 256$ subbands, a warped FBE with $M = 64$ subbands, and a warped AR LDF of degree $P = 11$. The right plot magnifies the highlighted area of the left plot.

frequency independent factor such that the output power of each speech frame is the same as after processing with the uniform AS FB.

It can be seen that using the warped FBE gives a performance in terms of SII comparable to the use of the uniform AS FB, which in turn results in a better speech intelligibility than the frequency independent amplification. Informal listening tests, however, showed a preference when using the warped FBE over the uniform AS FB.

The mean SII after processing with the warped autoregressive (AR) filter approximation and $M = 64$ subbands (dashed line) is compared to the above mentioned algorithms. It can be observed that the AR LDF achieves almost the same performance as the warped FBE and outperforms the uniform AS FB with 256 subbands.

The processing with the uniform AS FB as proposed in [1] has an algorithmic delay of 159 samples due to frame processing and overlap-add method. Using the warped FBE reduces the algorithmic delay to 95 samples whereas the AR LDF has an algorithmic delay between 0 and 2 samples due to the changing phase response of the AR filter.

4 Conclusions

A new algorithm for near end listening enhancement by using a non-uniform low delay filter-bank is proposed. The utilized frequency warped filter-bank equalizer performs time-domain filtering with coefficients adapted in the frequency domain. This allows for a processing with non-uniform spectral resolution and low signal delay.

The calculation of the spectral weights is done for approximately Bark-scaled frequency bands to account for the non-uniform frequency resolution of the human ear. In the process, the average speech spectrum of the received signal from the far end speaker is raised over the average noise spectrum to achieve an improved speech intelligibility for the near end speaker. The calculation of the needed power spectral densities of the far end speech and background noise is done with respect to the non-uniform time-frequency resolution of the warped filter-bank.

The instrumental evaluation by means of the Speech Intelligibility Index has shown that the new system achieves the same speech quality with lower signal delay than a system based on a uniform DFT AS FB.

It is also possible to approximate the warped time-domain filter of the filter-bank equalizer by a warped autoregressive filter of lower degree. This allows to reduce the signal delay to only a few samples without impairments for the speech intelligibility.

References

[1] B. Sauert and P. Vary. Near end listening enhancement: Speech intelligibility improvement in noisy environments. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 493–496, May 2006.

[2] E. Zwicker and H. Fastl. *Psychoacoustics: Facts and Models*. Springer, Berlin, 2nd edition, 1999.

[3] J. O. Smith and J. S. Abel. Bark and ERB bilinear transforms. *IEEE Transactions on Speech and Audio Processing*, 7(6):697–708, November 1999.

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

[5] H. W. Löllmann and P. Vary. Uniform and warped low delay filter-banks for speech enhancement. *Speech Communication*, 49:574–587, July 2007. Special issue on Speech Enhancement.

[6] H. W. Löllmann and P. Vary. Low Delay Filter-Banks for Speech and Audio Processing. In E. Hänsler and G. Schmidt, editors, *Speech and Audio Processing in Adverse Environments*, chapter 2, pages 13–61. Springer, Berlin, New York, 2008.

[7] H. W. Löllmann and P. Vary. Parametric Phase Equalizers for Warped Filter-Banks. In *Proceedings of European Signal Processing Conference (EUSIPCO)*, Florence, Italy, September 2006.

[8] American National Standard. Methods for the Calculation of the Speech Intelligibility Index. ANSI S3.5-1997, 1997.