Frequency Domain De-Essing for Hands-free Applications

Klaus Linhard, Philipp Bulling, Arthur Wolf

Daimler AG, D-89081 Ulm, Germany, Email: {klaus.linhard, philipp.bulling, arthur.wolf}@daimler.com

Abstract

A de-essing algorithm for applications in noisy environments, such as hands-free systems or in-car communication systems in cars, is presented. De-essing is a technique to reduce sibilance in speech or vocal recordings. In human voice, sibilance is caused by sibilant consonants, which belong to the fricatives. Sibilant consonants mainly consist of frequencies in the range of 2 kHz to 8 kHz and they are perceived as a hissing sound. For speech recordings in cars, usually the microphone frequency response already reduces the noisy lower frequencies, i.e. the higher frequencies and thus the sibilants are emphasized. The proposed de-esser is capable of detecting sibilants and damping the corresponding frequency bands. The realization is based on a Discrete Fourier Transform filterbank. The algorithm consists of three reduction filters, controlled by relative thresholds. The first filter is a notch filter. Its center frequency is tuned automatically to the frequency of the maximum level of a sharp sibilant. The second filter is a broadband bandstop filter, used for higher frequencies. A third filter interpolates the range between the minima of the first two filters. Speech recordings in German language, with sibilant consonants such as “s”, “ss”, “sch”, “z” and “tz”, are used for evaluation.

Introduction

De-essing is a technique to reduce sibilant sounds in speech signals [1]. Algorithms for de-essing are dynamic processors that damp frequency bands with hissing sounds. The basis is a dynamic range compressor, as for example presented in [2]. A compressor maps the level of the input signal to a smaller output level, if a fixed threshold is exceeded. This can be either applied to the broadband signal, or to individual frequency bands. The latter is often called multi-band compression. Since sibilants also occur at small signal levels, in contrast to dynamic range compression a de-esser works independently of the absolute signal level. Therefore, the level of the critical hissing frequency bands is compared to the average signal level [3]. If this relation exceeds a relative threshold, the corresponding frequency bands are damped. One or more bandpass filters in a side-chain are used to calculate the level of the sibilants. Most existing approaches use time domain FIR or IIR filters to process the signal [4]. In hands-free or in-car communication applications, signals are often processed in the frequency domain by means of filterbanks. For this purpose a novel frequency domain de-esser is presented in this paper. The proposed de-esser is adaptive, meaning that sibilant frequencies are detected automatically. By doing so, real-time processing of various different speakers is possible.

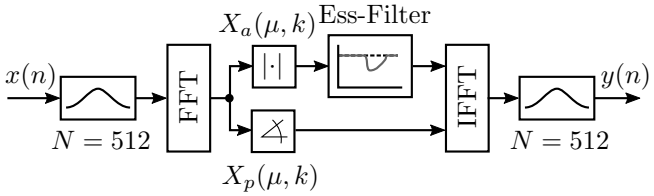

Figure 1: De-esser within a DFT filterbank structure.

Frequency Domain De-Essing

The realization is based on a Discrete Fourier Transform (DFT) filterbank. Three reduction filters are controlled by relative thresholds. The first reduction filter is a notch filter with its center frequency tuned to the frequency of the maximum level of a sharp sibilant. The second reduction filter is a broadband bandstop filter, used for higher frequencies. A third filter is used to interpolate the range between the minima of the first two filters. The motivation for this interpolation is to create a smoother shape, resulting in a more pleasant sound.

DFT-Filterbank

The input signal $x(n)$ is segmented into blocks of size $N = 512$ samples, with 3/4 overlap or frameshift $L = 128$. The sampling rate is $f_s = 32\,\text{kHz}$. n denotes the discrete time index. The resulting block samples are transformed to the frequency domain using Fast Fourier Transform (FFT) of length N. For windowing, the square root of a Hanning window is used. The calculated “ess”-filter is applied only on the magnitude of the frequency samples $X_a(\mu,k)$, while the phase $X_p(\mu,k)$ is not processed. μ is the discrete frequency index and $k = n/L$ is the subsampled block index. The output signal $y(n)$ is obtained by weighting the Inverse Fourier Transform (IFFT) with the square root of the Hanning window and addition of the overlapping segments. The corresponding block diagram is shown in fig. 1.

Frequency Analysis and Filtering

For analysis, frequencies below 200 Hz are removed with a highpass filter (HP). Two bandpass filters are used to detect the sibilants. The first bandpass filter (BP1) creates a search range for sharp “ess”-sounds. The notch filter with tunable center frequency will be placed within this range later, in order to reduce the narrowband “ess”-part. The second bandpass filter (BP2) is applied to the higher frequency range. In case of higher power level within this range, a bandstop filter (hiss filter) will be applied. The minima of the notch and the hiss filter are

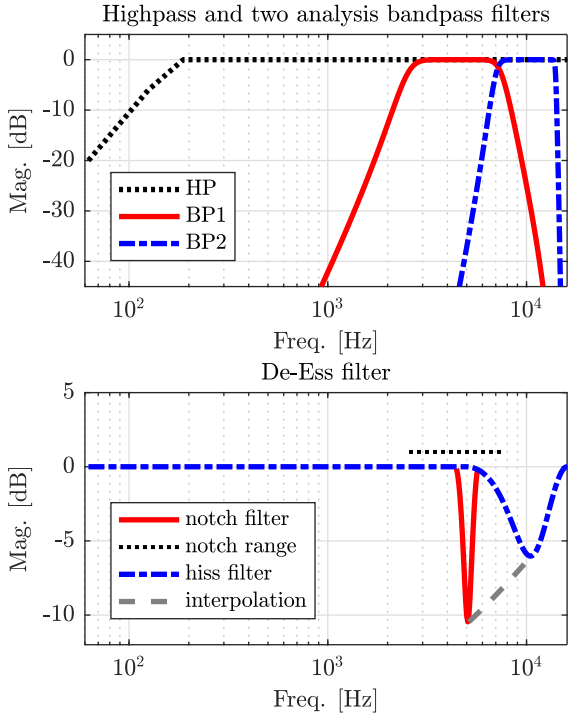

Figure 2: Top: Analysis filter. Bottom: Notch, hiss and interpolation filter.

interpolated to provide the final filter. The magnitude responses of the filters are shown in fig. 2. Both notch and hiss filter are deduced from a Hanning window.

Normalizing the Input Signal Level

The power within the frequency range of the hiss filter is denoted with $P_{\text{hiss}}(k)$

$$P_{\text{hiss}}(k) = \sum_{\mu=0}^{N-1} |X_{\text{BP2}}(\mu, k)|^2 , \tag{1}$$

where $X_{\text{BP2}}(\mu, k)$ is the signal after weighting with BP2. In the frequency range of the notch filter, the frequency bins are smoothed

$$X_{\text{BP1,smo}}(\mu, k) = \frac{1}{a} \cdot \sum_{i=0}^{a-1} |X_{\text{BP1}}(\mu - i, k)| . \tag{2}$$

Since smoothing is performed over the frequency bins here, this filter can also be realized as non-causal zero-phase filter. After smoothing, the frequency $f_0(k)$ with maximum power level is determined. Later, the notch filter is positioned at $f_0(k)$. The maximum level is multiplied with the number of coefficients a that are used for smoothing, resulting in power $P_{\text{notch}}(k)$

$$P_{\text{notch}}(k) = a \cdot |\max \{X_{\text{BP1,smo}}(\mu, k)\}|^2 . \tag{3}$$

In the following $a = 7$ bins are used, corresponding to a bandwidth of approximately 438 Hz.

Both power values are normalized with the highpass filtered total signal power $P(k)$

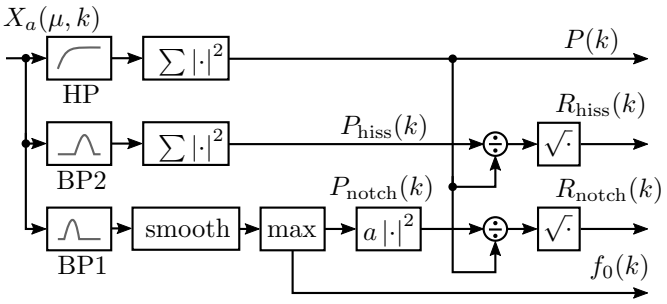

Figure 3: Analysis of input signal and power normalization.

$$P(k) = \sum_{\mu=0}^{N-1} |X_{\text{HP}}(\mu, k)|^2 . \tag{4}$$

To reduce the dynamic range of the power ratios, the square root is taken, resulting in the ratios $R_{\text{hiss}}(k)$ and $R_{\text{notch}}(k)$

$$R_{\text{hiss}}(k) = \sqrt{\frac{P_{\text{hiss}}(k)}{P(k)}} \tag{5}$$

$$R_{\text{notch}}(k) = \sqrt{\frac{P_{\text{notch}}(k)}{P(k)}} . \tag{6}$$

The principle is shown schematically in fig. 3.

Both ratios, R_{hiss} and R_{notch}, are smoothed with a peak detector, as explained in [2]. With attack time AT and release time RT, a fast onset and a slow drop off is realized. The peak value for every block can then be calculated using the following equation for both $R_{\text{hiss}}(k)$ and $R_{\text{notch}}(k)$

$$\overline{R(k)} = (1 - \tau(k)) \cdot \overline{R(k-1)} + \tau(k) \cdot R(k). \tag{7}$$

$\overline{(\cdot)}$ denotes a smoothed value, $\tau(k)$ is the time dependent smoothing constant

$$\tau(k) = \begin{cases} AT & \text{if } R(k) > \overline{R(k-1)} \\ RT & \text{else.} \end{cases} \tag{8}$$

For the notch filter, the values are set to $AT_{\text{notch}} = 0.7$ and $RT_{\text{notch}} = 0.4$. The values for the hiss filter are $AT_{\text{hiss}} = 0.7$ and $RT_{\text{hiss}} = 0.1$.

Compressor Characteristics

After peak detection, both power ratios $\overline{R_{\text{hiss}}(k)}$ and $\overline{R_{\text{notch}}(k)}$ are compared to the threshold T_1. For both filters, the threshold is set to -10 dB. If the threshold is passed, the corresponding filter is applied. A second threshold T_2 is reached at -6 dB. If T_2 is passed, a second set of notch or hiss filters with a stronger damping is used. In many cases only one filter set may be sufficient. Nevertheless, the extension to two filter sets offers a more flexible reduction.

As in conventional dynamic range compressors, gain characteristics in decibel can be provided if reduction

is denoted as a negative gain [2]. The normalized and smoothed power $\overline{R(k)}$ is used as input to the characteristics. The compression curve is shown in fig. 4. Below threshold T_1, the gain is $G = 0\,\mathrm{dB}$, i. e. no damping is applied. If $\overline{R(k)}$ reaches T_1, reduction starts and for $\overline{R(k)} = T_2$ the gain is $G = -3\,\mathrm{dB}$. At $\overline{R(k)} = 0\,\mathrm{dB}$ the gain is $G = -10\,\mathrm{dB}$ in case of the notch filter and $G = -6\,\mathrm{dB}$ in case of the hiss filter. Gain values in between the thresholds are obtained using linear interpolation.

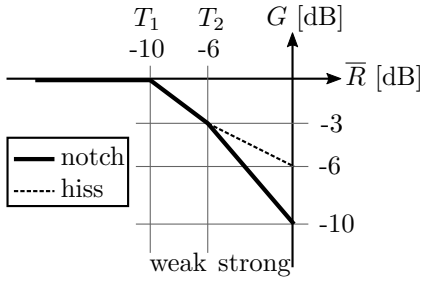

Figure 4: Gain characteristics.

Fig. 5 shows the filter curves for input $\overline{R(k)} = T_2$ and $\overline{R(k)} = 0\,\mathrm{dB}$. The range $T_1 < \overline{R(k)} < T_2$ is declared as "weak" and $T_2 < \overline{R(k)} < 0\,\mathrm{dB}$ as "strong" in sense of the compression (reduction).

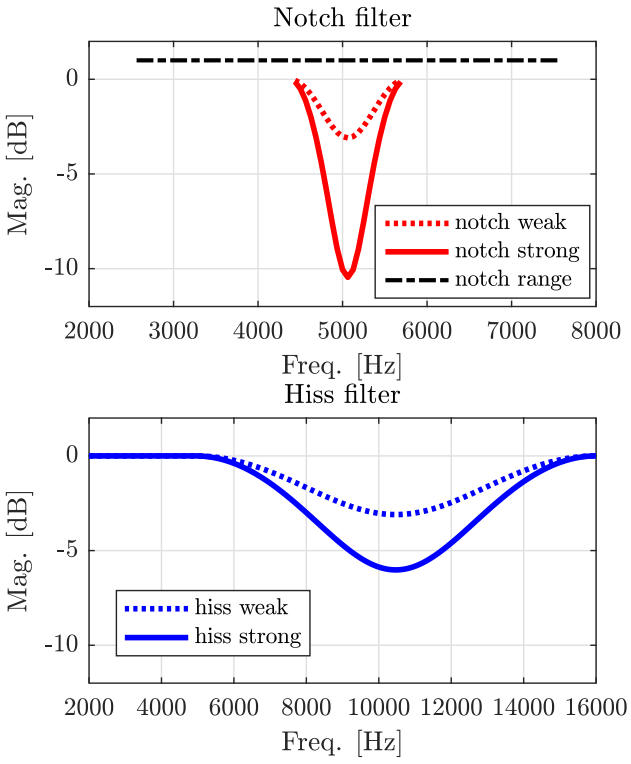

Figure 5: Top: Notch filter, weak and strong. Bottom: Hiss filter, weak and strong.

After the determination of notch and hiss filters, an interpolation line is calculated between the two minima, which finally results in the overall filter. This interpolation can only be done if there are two minima, otherwise it is omitted. An example is already shown in fig. 2 for

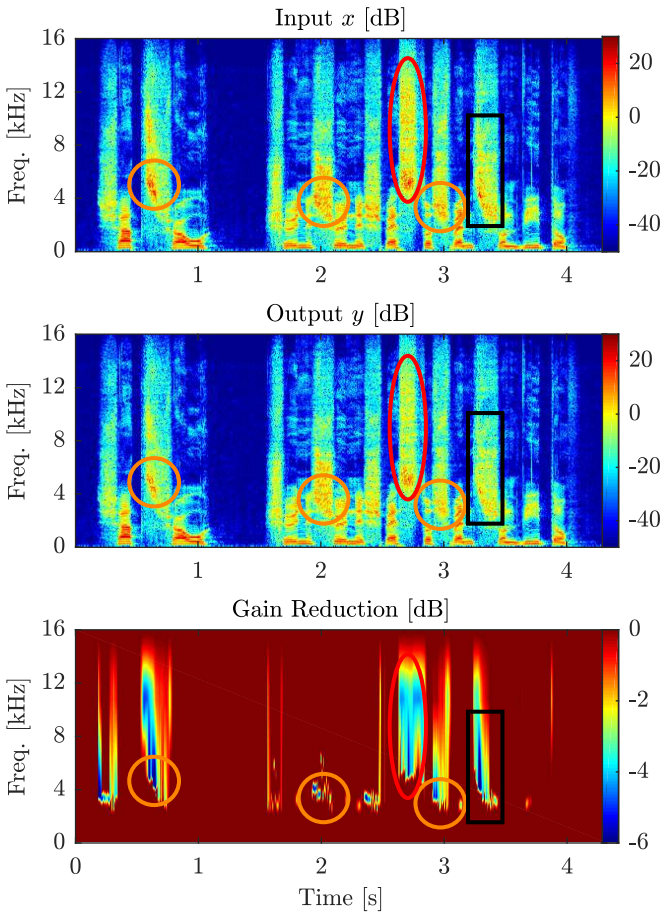

Figure 6: Spectrum of input and output signal, as well as frequency selective gain reduction.

the case of a notch minimum of -10 dB and a hiss minimum of -6 dB.

Simulation Results and Implementation

The German sentence "Schatz schau: Schönis schöne Schwester schwänzt schon wieder" from [5] is used to show the results of the algorithm. The sentence is spoken by a male speaker with clearly perceivable hiss sound. The microphone has a linear frequency response. However, the low frequencies are reduced, to simulate a moderate hands-free response. This filtering will emphasize the hiss effect, but reducing the lower frequency content is standard for hands-free applications. The resulting frequency response lies within the VDA tolerance band, given by [6]. The sentence contains seven "sch" phonems, one "s" phoneme in "Schwester" and one "zt" in "schwänzt". In fig. 6 in the upper plots the spectrograms of input and output are shown. Sibilant consonants are marked as follows:

- Orange circle: "sch"
- Red ellipse: "s"
- Black rectangle: "zt"

Looking at these sibilants, both narrowband and broadband areas can be seen. The lower plot of fig. 6 illustrates the gain reduction, performed by the algorithm.

Notch center frequencies and the hiss range for weak and strong sibilance, as well as the interpolation range (between notch and hiss) can be seen. All manually marked areas are detected and additionally some more critical frequencies are found and submitted to the filtering. In this example, dynamic filtering results in a reduction of the total signal level of about 1 dBA.

Real-time tests in a demonstration car showed that the de-esser performs reliable for various speakers and also in presence of background noise. Both hands-free systems and in-car communication systems benefit from reduced sibilants. Nevertheless, the sound of the processed speech remains natural. In addition, in-car communication systems have to deal with howling caused by the closed electro-acoustic loop. A positive side-effect is that the de-esser also suppresses these howling artifacts to a certain extend.

Conclusions

A de-esser was realized as dynamic compressor with parametric equalizing in frequency domain. The powers of the bandpass filtered input signal are normalized and thus they are dynamically adjusted. A notch filter follows with its center frequency the maximum power within a defined search range. The depth of the notch is adjusted depending on the power. A second broadband filter with a controlled minimum is used for higher frequencies. The realization based on a DFT filterbank allows an easy combination with other hands-free algorithms also using a DFT filterbank.

The shown analysis filters are given only for formal reasons and need not to be implemented in practice because calculation can directly be done on the frequency bins. The given extension with two thresholds may be simplified in many cases to the use of only one threshold. In summary, the approach is very flexible with only moderate implementation effort.

References

[1] M. Senior, "Techniques for vocal de-essing." `http://www.soundonsound.com/techniques/techniques-vocal-de-essing`, May 2009.

[2] U. Zölzer, X. Amatriain, and D. Arfib, *DAFX: digital audio effects*, vol. 1. Wiley Online Library, 2011.

[3] R. Jeffs, S. Holden, and D. Bohn, "Dynamics processors - technology & application tips," *Rane Corporation*, 2005.

[4] M. Wolters, M. Sapp, and J. Becker-Schweitzer, "Adaptive algorithm for detecting and reducing sibilants in recorded speech," in *Audio Engineering Society Convention 104*, May 1998.

[5] L. Hyttynen, "Zungenbrecher, die mit s beginnen." `http://www.bookanddrink.com/kinder/zungenbrecher/deutsche/s.htm`. Accessed: January 2017.

[6] ITU-T, Recommendation P.1110, "Wideband hands-free communication in motor vehicles," Jan. 2015.